NAVAL RESEARCH LOGISTICS QUARTERLY

SEPTEMBER 1971
VOL. 18, NO. 3

OFFICE OF NAVAL RESEARCH

NAVSO P-1278

NAVAL RESEARCH LOGISTICS QUARTERLY

The Naval Research Logistics Quarterly is devoted to the dissemination of scientific information in logistics and will publish research and expository papers, including those in certain areas of mathematics, statistics, and economics, relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Information for Contributors is indicated on inside back cover.

The Naval Research Logistics Quarterly is published by the Office of Naval Research in the months of March, June, September, and December and can be purchased from the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402. Subscription Price: $5.50 a year in the U.S. and Canada, $7.00 elsewhere. Cost of individual issues may be obtained from the Superintendent of Documents.

Issuance of this periodical approved in accordance with Department of the Navy Publications and Printing Regulations, NAVEXOS P-35

ELIMINATION METHODS IN THE $m \times n$ SEQUENCING PROBLEM †

Wlodzimierz Szwarc

University of Wisconsin—Milwaukee
Milwaukee, Wisconsin

ABSTRACT

This paper considers elimination methods in solving the sequencing problem where no passing is permitted. An elimination method consists of reducing (according to some criterion) the initial set of n! solutions to a smaller set. A crucial question arises as to whether this reduced set contains an optimal solution. The answer is affirmative if this elimination criterion implies condition (3).

1. INTRODUCTION

Consider the following $m \times n$ sequencing problem. There are given m machines $M_1, M_2, \ldots, M_m$ and n jobs $a, b, \ldots, v$. Each job is to be processed on the machines in the same order $M_1, M_2, \ldots, M_m$. We also assume that the processing order of the jobs is the same on each machine. Each machine starts operating a job as soon as it is free and the job is available for processing.

Given the processing times for each job on each machine find the operation order (sequence) of the jobs with the minimal total elapsed time (from a fixed starting moment on machine M_1 till the moment when machine M_m finishes the last job of the sequence). Such a sequence will be called optimal sequence or optimal solution. We will use the following notations:*

$a_k, b_k, \ldots, v_k$ are the processing times of jobs $a, b, \ldots, v$ on machine M_k.

σ, π, ϵ, are sequences (possibly empty) of jobs, σa is a sequence consisting of a and jobs of σ (other than a) preceding job a.

Let $t(\sigma a, k)$ be the completion time on machine M_k of all jobs of sequence σa (counted from a moment when machine M_1 starts operating the jobs of σa). Then

$$t(\sigma a, k) = \max\,[t(\sigma a, k-1), \quad t(\sigma, k)] + a_k \tag{1}$$

with

$$t(\phi, k) = 0 = t(\sigma, 0).$$

Then $t(\sigma, m)$ is the total elapsed processing time for σ on machines $M_1, \ldots, M_m$.

This paper presents a method of solving the $m \times n$ sequencing problem which reduces the initial set of $n!$ sequences to a smaller subset called reduced set.

Let a and b be two jobs and σ a sequence of jobs other than a and b.

An elimination rule reads: If some condition (which will be called elimination criterion) holds,

---

†This work was completed when the author was a member of the faculty of the Carnegie-Mellon University, Pittsburgh, Pennsylvania.

*The notations are essentially the same as those in [1].

then eliminate all sequences beginning with σb. A crucial question arises as to whether the reduced set always contains an optimal solution. To answer that question, first introduce some notations.

Let π', π'' be arbitrary (even empty) sequences of jobs satisfying conditions

$$\left.\begin{aligned} &\pi' \cap \pi'' = \phi, \qquad (\pi' \cup \pi'') \cap \sigma ab = \phi \\ &(\pi' \cup \pi'') \cup \sigma ab = \{a, b, \ldots, v\} \end{aligned}\right\}. \tag{2}$$

The following is true:

If for any π' and π'' satisfying (2) an elimination criterion implies

$$t(\sigma ab\pi'\pi'', m) \leq t(\sigma b\pi' a\pi'', m) \tag{3}$$

then the best solution of the reduced set is the optimal solution of the $m \times n$ sequencing problem.

Let

$$\Delta_k \overset{\text{df}}{=} t(\sigma ab, k) - t(\sigma b, k). \tag{4}$$

This paper proposes the following elimination criterion*

$$\Delta_{k-1} \leq \Delta_k \leq a_k, \qquad \text{for} \qquad k = 2, \ldots, m. \tag{IV}$$

Section 2 contains a proof that criterion (IV) implies (3). This means that the method based on (IV) will always produce an optimal solution of the problem.

REMARK 1: One may look at this problem in a different way by fixing the end moment of the operation program and then look for a processing order whose time is running in the opposite direction. Then instead of (1) we can write

$$\bar{t}(a\sigma', k') = \max\left[\bar{t}(a\sigma', k' - 1), t(\sigma', k')\right] + a_{k'}, \tag{1'}$$

where σ' is a permutation of jobs of σ ordered in the opposite direction and $k' = m - k + 1$. It is easy to see that

$$t(\sigma a, m) = \bar{t}(a\sigma', 1).$$

Consider criterion (IV') which is symmetrical to (IV). All we have to do is arrange the machines and jobs in the reverse order and construct for this case condition (IV). We define a new elimination rule: if (IV') holds then remove all sequences ending with $b\sigma$.†

The proof that the reduced set obtained by this rule will contain an optimal sequence is symmetrical to the proof of the theorem (IV) implies (3). Section 4 examines the properties of the proposed elimination rule and presents an outline of the solution procedure illustrated on two numerical examples. The proposed elimination criteria imply several simple rules which are of considerable convenience in the solution process. These rules offer a deeper insight into this problem. So, for instance, if there is no jobs c such that for all k: $c_1 \leq c_k$ or $c_m \leq c_k$, then no single sequence could be eliminated. In this case the elimination method offers no advantage since we would have to consider all $n!$ sequences.

Section 3 examines the properties of all known elimination criteria as well as the connections between them. Smith and Dudek gave in the errata to [5] the following criterion:

---

*The elimination rule reads: if (IV) holds then eliminate all sequences of the form $\sigma b \ldots$.

†The elements of σ are in (IV') arranged in an opposite direction.

(II)
$$\left.\begin{aligned} \Delta_{k-1} &\leq a_k \\ t(\sigma a, k-1) &\leq t(\sigma b, k-1) \end{aligned}\right\} k = 2, \ldots, m.$$

We show in section 3 that (IV) is a weaker condition than (II), thereby proving implicitly that (II) also implies (3).* Since rule (IV) is weaker than (II), it will *eliminate more sequences* (or at least as many) than (II).

According to Proposition 3, no other (of the known) criterion implies (3). One may ask whether one could develop a criterion weaker than (IV) and which implies (3) (see Remark 5, section 3).

Another question of interest is whether to explore the possibility of applying the elimination methods to find approximate solutions to the sequencing problem. For instance, one may look for elimination rules which produce a reduced set consisting of good solutions or alternatively whose best sequence is a good solution.

2. PROOF OF THE MAIN THEOREM.

Let $\pi = (p^1, p^2, \ldots, p^r)$ be an arbitrary sequence of jobs, satisfying condition

$$\pi \subset \{a, b \ldots, v\} - \{\sigma ab\}.$$

We will prove the following:

LEMMA 1:

$$(IV) \Rightarrow t(\sigma ab\pi, k) - t(\sigma b\pi, k) \leq t(\sigma ab, k) - t(\sigma b, k) \qquad \text{for} \qquad k = 1, 2, \ldots, m.$$

PROOF: *Step 1.* Let $r = 1$. Then $\pi = p^1$. Proof by induction. The theorem is true for $s = 1$ since $t(\sigma abp^1, 1) - t(\sigma bp^1, 1) = a_1 = t(\sigma ab, 1) - t(\sigma b, 1)$. Suppose the theorem is true for $s = k - 1$. We will prove it for $s = k$. Consider

$$t(\sigma abp^1, k) - t(\sigma bp^1, k) = \max\,[t(\sigma abp^1, k-1), t(\sigma ab, k)] + p^1_k - \max\,[t(\sigma bp^1, k-1), t(\sigma b, k] - p^1_k)\dagger$$

$$\leq \max\,[t(\sigma abp^1, k-1) - t(\sigma bp^1, k-1), t(\sigma ab, k) - t(\sigma b, k)] \text{ (inductive assumption)}$$

$$\leq \max\,[t(\sigma ab, k-1) - t(\sigma b, k-1), t(\sigma ab, k) - t(\sigma b, k)] \text{ (condition (IV))}$$

$$\leq t(\sigma ab, k) - t(\sigma b, k), \qquad \text{QED.}$$

Step 2. Having already proved the theorem for $r = 1$, set $r = 2$. Then $\pi = (p^1, p^2)$. For $s = 1$, the theorem holds. Assuming it is true for $s = k - 1$, we will prove it for $s = k$. As in step 1, by using the footnote, we get

$$t(\sigma abp^1p^2, k) - t(\sigma bp^1p^2, k) \leq \max\,[t(\sigma abp^1p^2, k-1) - t(\sigma bp^1p^2, k-1), t(\sigma abp^1, k) - t(\sigma bp^1, k)]$$

$$\text{(inductive assumption)} \leq \max\,[t(\sigma ab, k-1) - t(\sigma b, k-1), t(\sigma abp^1, k) - t(\sigma bp^1, k)] \text{ (result from step 1)} \leq \max\,[t(\sigma ab, k-1), -t(\sigma b, k-1), t(\sigma ab, k) - t(\sigma b, k] \text{ (by (IV))} \leq t(\sigma ab, k) - (\sigma b, k), \qquad \text{QED.}$$

Performing steps 3, 4, . . ., r we obtain the necessary result.

---

*It is not clear whether the text on pp. 75–76 of [5] is a complete proof that (II) implies (3). Note that $K(\sigma ab, k) \leq K(\sigma b, k)$ (see p. 72) in [5] is equivalent to (V), i.e., to: $\Delta_{k-1} \leq \Delta_k$, for all k. However, (V) $\not\Rightarrow$ (3) (see Proposition 3 in section 3).

† $\max\,(A, B) - \max\,(C, D) \leq \max\,(A - C, B - D)$.

LEMMA 2: † Let ϵ and ϵ' be two different permutations of the same jobs and let

$$\pi = (p^1, p^2, \ldots, p^r)$$

be an arbitrary sequence of jobs satisfying

$$\pi \subset \{a, b, \ldots, v\} - \epsilon.$$

Then for $k = 1, \ldots, m$

$$t(\epsilon, k) \leqslant t(\epsilon', k) \Rightarrow t(\epsilon\pi, k) \leqslant t(\epsilon'\pi, k).$$

PROOF: *Step 1.* $r = 1$, then $\pi = (p^1)$. The theorem is true for $s = 1$ since $t(\epsilon, 1) = t(\epsilon', 1)$ and $t(\epsilon p^1, 1) = t(\epsilon' p^1, 1)$. Assuming the theorem to be true for $s = k-1$, we will prove it for $s = k$. According to (1),

$$t(\epsilon p^1, k) \leqslant t(\epsilon' p^1, k) \Longleftrightarrow \max\,[t(\epsilon, k), t(\epsilon p^1, k-1)] + p^1_k \leqslant \max\,[t(\epsilon', k), t(\epsilon' p^1, k-1)] + p^1_k$$

which holds since $t(\epsilon, k) \leqslant t(\epsilon', k)$ (assumption), and $t(\epsilon p^1, k-1) \leqslant t(\epsilon' p^1, k-1)$ (inductive assumption).

Step 2. $r = 2$, $\pi = (p^1, p^2)$. We have proved that $t(\epsilon p^1, k) \leqslant t(\epsilon' p^1, k)$. Regard $\bar{\epsilon} = \epsilon p^1$, $\bar{\epsilon}' = \epsilon' p^1$ and $\bar{\pi} = p^2$. Using the result from Step 1, we have $t(\epsilon p^1, k) \leqslant t(\epsilon' p^1, k) \Rightarrow t[(\epsilon p^1)p^2, k] \leqslant t[(\epsilon' p^1)p^2, k]$, but $t(\leqslant, k) \leqslant t(\epsilon', k) \Rightarrow t(\epsilon p^1, k) \leqslant t(\epsilon' p^1, k)$ (Step 1); hence $t(\epsilon, k) \leqslant t(\epsilon', k) \Rightarrow t(\epsilon p^1 p^2, k) \leqslant t(\epsilon' p^1 p^2, k)$, QED.

Conducting Steps 3, 4, . . ., r we obtain the necessary result. We are ready to prove the following

MAIN THEOREM: (IV) $\Rightarrow$ (3) for all $k = 1, \ldots, m$.

PROOF: Setting π' instead of π in lemma 1, we get

$$\text{(IV)} \Rightarrow t(\sigma ab\pi', k) - t(\sigma b\pi', k) \leqslant \Delta_k, \quad k = 1, \ldots, m.$$

Since $\Delta_k \leqslant a_k$ (by (IV)) we have

$$t(\sigma ab\pi', k) \leqslant t(\sigma b\pi', k) + a_k \qquad k = 1, \ldots, m.$$

This together with (1) implies for all $k = 1, \ldots, m$

$$t(\sigma ab\pi', k) \leqslant t(\sigma b\pi', k) + a_k \leqslant t(\sigma b\pi' a, k).$$

Therefore, for all k

$$t(\sigma ab\pi', k) \leqslant t(\sigma b\pi' a, k).$$

Applying Lemma 2 (where $\epsilon' = \sigma b\pi' a$), we have

$$t(\sigma ab\pi'\pi'', k) \leqslant t(\sigma b\pi' a\pi'', k) \text{ for } k = 1, \ldots, m.$$

Hence for $k = m$

$$t(\sigma ab\pi'\pi'', m) \leqslant t(\sigma b\pi' a\pi'', m), \text{ QED.}$$

3. CONNECTION BETWEEN THE ELIMINATION CRITERIA.

Consider all known elimination criteria:*

1° [2].

† This is a known theorem, see [3] and [5].

*Criteria (I), (II), and (V) are written in an equivalent form.

(I) $$t(\sigma ab, k) \leq t(\sigma ba, k), \qquad k=2, \ldots, m.$$

2° [5] (errata).

(II) $$\left.\begin{aligned} \Delta_{k-1} &\leq a_k \\ t(\sigma a, k-1) &\leq t(\sigma b, k-1) \end{aligned}\right\} k=2, \ldots, m.$$

3° [1].

(III) $$\Delta_k \leq a_k \qquad k=2, \ldots, m.$$

Let us include also Condition (3) from [5] (page 72).

4° [5].

(V) $$\Delta_{k-1} \leq a_k \qquad k=2, \ldots, m.$$

Our goal is to examine and to establish connections between all the five criteria. First we will introduce some definitions.

Let $t_{\sigma, k}$ be the sum of nominal operation times on machine M_k of all the jobs belonging to sequence σ.*

By $Q(\sigma, k)$ denote the total waiting time on machine M_k until the moment when the last job of σ is processed. Then (see (1))

(5) $$Q(\sigma, k) = t(\sigma, k) - t_{\sigma, k}.$$

Smith and Dudek introduced in [5] a term

$$K(\sigma a, k) = t_{\sigma a, k-1} - t_{\sigma, k} + Q(\sigma a, k-1),$$

which in view of (5) can be rewritten in the following equivalent form

(6) $$K(\sigma a, k) = t(\sigma a, k-1) - t_{\sigma, k}.$$

We will show:

LEMMA 3†.

(7) $$Q(\sigma a, k) = \max [Q(\sigma, k), \qquad K(\sigma a, k)].$$

PROOF: By subtracting $t_{\sigma a, k} = t_{\sigma, k} + a_k$ from both sides of (1) we get

$$t(\sigma a, k) - t_{\sigma a, k} = \max [t(\sigma, k) - t_{\sigma, k}, \qquad t(\sigma a, k-1) - t_{\sigma, k}],$$

which is (7) (according to (5) and (6)).

The authors of [5] presented in the errata the following criterion.

(*) $$\max [K(\sigma a, k), K(\sigma ab, k)] \leq K(\sigma b, k), \qquad k=2, \ldots, m.$$

Using (6) one can see that (*) is equivalent to (II).

LEMMA 4: (II) $\Rightarrow$ (III).

PROOF: (II) $\Rightarrow$ (*) $\Rightarrow$ $\max [Q(\sigma, k), K(\sigma a, k), K(\sigma ab, k)] \leq \max [Q(\sigma, k), K(\sigma b, k)] \leq \Longleftrightarrow$ (by (7)) $\Longleftrightarrow$ $\max [Q(\sigma a, k), K(\sigma ab, k)] \leq Q(\sigma b, k) \Longleftrightarrow$ (by (7)) $\Longleftrightarrow Q(\sigma ab, k) \leq Q(\sigma b, k) \Longleftrightarrow$ (by (5)) $\Longleftrightarrow t(\sigma ab, k) - t_{\sigma ab, k} \leq t(\sigma b, k) - t_{\sigma b, k} \Longleftrightarrow t(\sigma ab, k) - t(\sigma b, k) \leq a_k$,

(since $t_{\sigma ab, k} = t_{\sigma b, k} + a_k$). QED.

---

*One may write $t_{\sigma, k}$ in form of a sum: $\sum_{c \in \sigma} t_{c, k}$, but then we would have to introduce a new symbol, $t_{c, k} \stackrel{\text{df}}{=} c_k$, for the operation time of job c on machine M_k.

†This result is known (see [5]).

LEMMA 5: (II) $\Rightarrow \Delta_{k-1} \leqslant \Delta_k$, $k = 2, \ldots, m$.

PROOF: According to (1) $t(\sigma b, k) - t(\sigma b, k-1) \geqslant b_k$.

There are two cases:

Case 1:

$$t(\sigma b, k) - t(\sigma b, k-1) = b_k. \tag{8}$$

We will examine $t(\sigma ab, k) - t(\sigma b, k-1)$. Adding $\Delta_k = t(\sigma ab, k) - t(\sigma b, k)$ to both sides of (8) we have

$$t(\sigma ab, k) - t(\sigma b, k-1) = \Delta_k + b_k. \tag{8'}$$

In view of (1)

$$t(\sigma ab, k) \geqslant t(\sigma ab, k-1) + b_k.$$

Adding $-t(\sigma b, k-1)$ to both sides of this inequality, we get

$$t(\sigma ab, k) - t(\sigma b, k-1) \geqslant \Delta_{k-1} + b_k. \tag{8''}$$

Comparing (8') and (8'') we obtain

$$\Delta_{k-1} \leqslant \Delta_k, \qquad \text{QED.}$$

Case 2:

$$t(\sigma b, k) - t(\sigma b, k-1) > b_k. \tag{9}$$

We will consider $t(\sigma b, k)$, $t(\sigma a, k)$, $t(\sigma ab, k)$, and Δ_k.

In view of (1) and (9)

$$t(\sigma b, k-1) < t(\sigma, k). \tag{9'}$$

Therefore

$$t(\sigma b, k) = t(\sigma, k) + b_k. \tag{9''}$$

(II) and (9') imply

$$t(\sigma a, k-1) \leqslant t(\sigma b, k-1) < t(\sigma, k).$$

Hence $t(\sigma a, k-1) < t(\sigma, k)$ and (by (1))

$$t(\sigma a, k) = t(\sigma, k) + a_k. \tag{9'''}$$

Consider

$$t(\sigma ab, k) = \max [t(\sigma a, k), t(\sigma ab, k-1)] + b_k = (\text{by (1), (4), (9''')})$$
$$= \max [t(\sigma, k) + a_k, t(\sigma b, k-1) + \Delta_{k-1}] + b_k.$$

Since $\Delta_{k-1} \leqslant a_k$, and $t(\sigma b, k-1) < t(\sigma, k)$ (by (II) and (9')), we have $t(\sigma ab, k) = [t(\sigma, k) + a_k] + b_k$. In view of (9'') we get

$$\Delta_k = t(\sigma ab, k) - t(\sigma b, k) = t(\sigma, k) + a_k + b_k - t(\sigma, k) - b_k = a_k.$$

Since (by (II)) $\Delta_{k-1} \leqslant a_k$ we have $\Delta_{k-1} \leqslant \Delta_k$, QED.

Let us show the following:

PROPOSITION 1: (II) $\Rightarrow$ (IV) $\Rightarrow$ (III) $\Rightarrow$ (I).

PROOF: From Lemmas 4 and 5 we have (II) $\Rightarrow$ (IV). Implication (IV) $\Rightarrow$ (III) follows directly from the definitions of (III) and (IV). According to (III) and (1) for each $k = 2, \ldots, m$

$$t(\sigma ab, k) \leq t(\sigma b, k) + a_k \leq t(\sigma ba, k) \Rightarrow t(\sigma ab, k) \leq t(\sigma ba, k)$$

which means that (III) $\Rightarrow$ (I), QED.

We will also prove:

PROPOSITION 2: (I) $\not\Rightarrow$ (III) $\not\Rightarrow$ (IV) $\not\Rightarrow$ (II).

PROOF: (IV) $\not\Rightarrow$ (II)

Consider a 3×3 example.

Example 1

	M_1	M_2	M_3
c	1^1	3^4	46^{50}
a	4^5	8^{13}	30^{80}
b	5^{10}	5^{18}	20^{100}

TABLE 1

	M_1	M_2	M_3
c	1^1	3^4	46^{50}
b	5^6	5^{11}	20^{70}

TABLE 1′

	M_1	M_2	M_3
c	1^1	3^4	46^{50}
a	4^5	8^{13}	30^{80}

TABLE 1″

REMARK 2: The lower numbers are operation times while the numbers above the operation times are the corresponding completion times found by formula (1). So, for instance $t(cab, 2) = 18$ (Table 1), $t(ca, 3) = 80$ (Table 1 or 1″), $t(cb, 1) = 6$ (Table 1′).

Here $\sigma = (c)$, and $\Delta_k = t(cab, k) - t(cb, k)$, $k = 1, 2, 3$. We have $\Delta_1 = 10 - 6 = 4$, $\Delta_2 = 18 - 11 = 7$, $\Delta_3 = 100 - 70 = 30$. As we can see, (IV) holds since $\Delta_1 < \Delta_2 < \Delta_3$, and $\Delta_1 = a_1 = 4$, $\Delta_2 < a_2 = 8$, $\Delta_3 = a_3 = 30$. Condition (II) does not hold since $t(ca, 2) - t(cb, 2) = 13 - 11 > 0$, QED.

(III) $\not\Rightarrow$ IV

Example 2 (from [6])

	M_1	M_2	M_3
p	1	5	6
q	3	5	2
r	5	2	3
s	2	7	5

	M_1	M_2	M_3
p	1^1	5^6	6^{12}
q	3^4	5^{11}	2^{14}
r	5^9	2^{13}	3^{17}

	M_1	M_2	M_3
p	1^1	5^6	6^{12}
r	5^6	2^8	3^{15}

Set $\sigma = (p)$, q as "a" r as "b", and $\Delta_k = t(pqr, k) - t(pr, k)$.

Then $\Delta_1 = 9 - 6 = 3$, $\Delta_2 = 13 - 8 = 5$, $\Delta_3 = 17 - 15 = 2$. $\Delta_1 = q_1 = 3$, $\Delta_2 = q_2 = 5$, $\Delta_3 = q_3 = 2$.

Condition (III) holds, whereas (IV) does not, since $\Delta_2 = 5 > \Delta_3 = 2$.

(I) $\not\Rightarrow$ (III)

Example 3 (from [4])

	M_1	M_2	M_3
a	3	22	2
b	22	20	20
c	20	14	18

	M_1	M_2	M_3
a	3^3	22^{25}	2^{27}
b	22^{25}	20^{45}	20^{65}

	M_1	M_2	M_3
b	22^{22}	20^{42}	20^{62}
a	3^{25}	22^{64}	2^{66}

Here $\sigma = \phi$. Condition (I) holds since $t(ab, k) \leq t(ba, k)$ for all k ($25 = 25$, $45 < 64$, $65 < 66$); however, (III) is not satisfied since $\Delta_3 = t(ab, 3) - t(b, 3) = 65 - 62 = 3 > a_3 = 2$, QED.

PROPOSITION 3: (I) $\not\Rightarrow$ (3), (III) $\not\Rightarrow$ (3), (V) $\not\Rightarrow$ (3).

PROOF: (I) $\not\Rightarrow$ (3).

Consider Example 3. Set $\sigma = \phi$. As already known, Condition (I) is satisfied for jobs a and b. We will show that (3) doesn't hold. Set $\pi' = (c)$, $\pi'' = \phi$. Here $m = 3$. Condition (3) is not satisfied since $t(abc, 3) = 83 > t(bca, 3) = 82$.

REMARK 3: According to elimination rule (I) we have to remove all sequences beginning with b. Sequence bca is, however, the only optimal solution (see [4]).

(III) $\not\Rightarrow$ (3).

Consider Example 2. As we have already shown (III) holds since $\Delta_k = t(pqr, k) - t(pr, k) \leq q_k$ for $k = 2, 3$. Set $\pi' = (s)$, $\pi'' = \phi$. Condition (3) is not met since $t(pqrs, 3) = 25 > t(prsq, 3) = 22$, QED.

REMARK 4: By removing all sequences which begin with pr (following the elimination rule based on (III)) we lose the only optimal sequence $prsq$.

REMARK 5: Consider Example 3. As known (3) as well as (III) do not hold. Consider condition

$$\text{(III*)} \qquad \Delta_{k-1} \leq \Delta_k \qquad k = 2, \ldots, m.$$

This example, the definition of (IV) as well as Propositions 1 and 2 show that

$$\text{(IV)} \Longleftrightarrow [\text{(III)} \cap \text{(III*)}] \Rightarrow \text{(3)}, \qquad \text{(III)} \not\Rightarrow \text{(III*)}, \ \text{(III*)} \not\Rightarrow \text{(III)}, \ \text{(III)} \not\Rightarrow \text{(3)}, \ \text{(III*)} \not\Rightarrow \text{(3)}.$$

Example 4

(V) $\not\Rightarrow$ (3)

	M_1	M_2	M_3
c	1	15	2
a	2	15	15
b	12	1	5
d	3	1	1

	M_1	M_2	M_3
c	1^1	15^{16}	2^{18}
a	2^3	15^{31}	15^{46}
b	12^{15}	1^{32}	5^{51}

	M_1	M_2	M_3
c	1^1	15^{16}	2^{18}
b	12^{13}	1^{17}	5^{23}

Assume $\sigma = (c)$, $\pi' = (d)$, $\pi'' = \phi$, $\Delta_k = t(cab, k) - t(cb, k)$. Criterion (V) holds since $\Delta_1 = 2 < a_2 = 15$, $\Delta_2 = 15 = a_3 = 15$; however, (3) does not hold for we have

$$t(cabd, 3) = 52 > t(cbda, 3) = 48, \text{ QED.}$$

REMARK 6: Notice that (V) $\not\Rightarrow$ (III). As we already know (V) holds whereas (III) does not ($\Delta_3 = 51 - 23 = 28 > a_3 = 15$). Therefore also (V) $\not\Rightarrow$ (IV).

COROLLARY 1: Propositions 1 and 2 imply that (IV) is a weaker condition than (II).

4. THE ELIMINATION METHOD.

We do not intend to present a step by step description of the solution procedure based on elimination criteria (IV) and (IV'). We offer instead an illustration of this procedure by solving in detail two numerical examples. I intentionally leave it for the reader to work out for himself an appropriate algorithm.* I do want, however, to call attention to some important properties of the elimination rule which may considerably improve the efficiency of the solution method.

*For guidance, see for instance a description of an algorithm given in [1].

Job b may be "removed" from the last position of sequence σb provided there exists a "remover," i.e., some job "a." Consider the elimination rules based on (IV)

$$\Delta_{k-1} \leq \Delta_k \leq a_k \qquad k=2, \ldots, m.$$

It follows from (1) that $\Delta_1 = a_1$. Hence $a_1 \leq \Delta_k \leq a_k$, which in turn implies: $a_1 \leq a_k$ for all $k=2, \ldots, m$.

This result leads to the following

COROLLARY 2: For a job a to be a remover (front remover) it is necessary that $a_1 \leq a_k$ for all $k=2, \ldots, m$. In view of Remark 1 (section 1) by using the symmetric elimination criterion, we will get a following

COROLLARY 3: For job a to be a back remover, it is necessary that $a_m \leq a_k$ for all $k=1, \ldots, m-1$. We will call jobs satisfying conditions of Corollary 2 (3) potential front (back) removers. By examining (IV) one may develop more useful rules, to mention two of them:

1° Instead of checking (IV) check the following condition equivalent to (IV)

$$\Delta_{k-1} \leq \Delta_k \leq a_k^* \qquad k=1, \ldots, m,$$

where $a_k^* = \min_{k \leq r \leq m} a_r$. The reason we introduced this condition is that it enables us to find out as soon as possible that (IV) *does no hold* (if this is the case). One can develop also an appropriate condition when checking (IV′).

2° Suppose we examine whether (IV) is true for $\sigma = \phi$ (i.e., when we try to remove jobs from the first position). Then a necessary condition for (IV) to hold is that $a_1 \leq b_1$.

Let us illustrate the elimination procedure by the following examples from section 3:

(A)

Example 1

	M_1	M_2	M_3
c	1	3	46
a	4	8	30
b	5	5	20

STAGE 0: *Find all potential front and back removers.*

Here all three jobs are potential front removers. Among the removers, choose a job with the minimal time on machine M_1, i.e., job c.

STAGE 1: *Elimination of jobs from the first position.*

Set $\sigma = \phi$ and job c as "a" and check whether (IV) is met when

(1) job a is "b," (2) job b is "b."

	M_1	M_2	M_3
c	1^1	3^4	46^{50}
a	4^5	8^{13}	30^{80}

a	4^4	8^{12}	30^{42}

	M_1	M_2	M_3
c	1^1	3^4	46^{50}
b	5^6	5^{11}	20^{70}

b	5^5	5^{10}	20^{30}

As seen, (IV) holds for both cases.

$$ad\ 1\colon \Delta_1 = 1 = c_1,\ \Delta_2 = 1 < 3 = c_2,\ \Delta_3 = 38 < 46 = c_3,\ 1 < 3 < 38.$$

$$ad\ 2\colon \Delta_1 = 1 = c_1,\ \Delta_2 = 1 < c_2,\ \Delta_3 = 40 < c_3,\ 1 = 1 < 40.$$

Hence, neither a nor b can occupy the first position, and we are to consider only sequences with c in the first place.

STAGE 2: *Elimination of jobs from the second position.*

We start with the only one presequence $\sigma = (c)$. Set job a as "a" and b as "b." Check whether (IV) is met. As we have already known (see Example 1 in section 3) (IV) holds. This means that we can remove all sequences beginning with cb. Hence, the only sequence to consider is cab where $t(cab, 3) = 100$. This sequence is optimal according to the Main Theorem.

(B) *Example 2*

	M_1	M_2	M_3
p	1	5	6
q	3	5	2
r	5	2	3
s	2	7	5

STAGE 0: *Find all potential removers.*

Jobs p and s are potential front removers while job q is a potential back remover.

STAGE 1: *Elimination of jobs from the first position.*

Set $\sigma = \phi$, job p as "a," and check whether (IV) holds when

(1) q is "b," (2) r is "b," (3) s is "b."

Since (IV) holds for all of the three cases, neither of the jobs q, r, s can occupy the first place. Hence we will consider only sequences beginning with p.

STAGE 1′: *Elimination of jobs from the last position.*

Set $\sigma' = \phi$, job q as "a," and check whether (IV′) holds when

(1) r is "b," (2) s is "b."

	M_3	M_2	M_1
q	2^2	5^7	3^{10}
r	3^5	2^9	5^{15}

	M_3	M_2	M_1
r	3^3	2^5	5^{10}

	M_3	M_2	M_1
q	2^2	5^7	3^{10}
s	5^7	7^{14}	2^{16}

	M_3	M_2	M_1
s	5^5	7^{12}	2^{14}

Condition (IV′) holds only for the second case since $\Delta_1' = 2 = q_3$, $\Delta_2' = 2 < q_2 = 5$, $\Delta_3' = 2 < q_1 = 3$, and $\Delta_1' = \Delta_2' = \Delta_3'$. This means that s cannot be the last element in the sequence.

STAGE 2: *Elimination of jobs from the second position.*

Since p is the only possible job in the first place, we consider only a presequence $\sigma = (p)$. We are left with job s as a potential front remover. Setting successively jobs q and r as jobs "b" and job s as "a," we can see that (IV) does not hold for either of these cases. Hence we cannot eliminate sequences which begin with pq, pr, (and ps as well).

STAGE 2′: *Elimination of jobs from the second last position.*

Since p must be the first job in the sequence (stage 1) and s cannot occupy the last place (Stage 1′), then only q and r can be the last jobs in the sequence. Since q is the only one potential remover, set $\sigma' = (r)$, q as "a," and s as "b."

	M_3	M_2	M_1
r	3^3	2^5	5^{10}
q	2^5	5^{10}	3^{13}
s	5^{10}	7^{17}	2^{19}

r	3^3	2^5	5^{10}
s	5^8	7^{15}	2^{17}

Condition (IV′) holds: $\Delta_1' = 2 = q_3$, $\Delta_2' = 2 < q_2 = 5$, $\Delta_3' = 2 < q_1 = 3$ and $\Delta_1' = \Delta_2' = \Delta_3'$. Hence we can eliminate all sequences ending with sr. Then the reduced set becomes (i.e., the set of sequences to consider) $prsq$, $psrq$, and $psqr$. Since $t(prsq, 3) = 22$, while $t(psrq, 3) = t(psqr, 3) = 23$, sequence $prsq$ is optimal.

REMARK 7: Note that by solving this example using elimination criterion (III) we would obtain a reduced set consisting of two sequences $pqsr$ and $pqrs$ where $t(pqsr, 3) = 26$, $t(pqrs, 3) = 25$ (see [6]), and neither of these sequences is even a good solution.

REFERENCES

[1] Bagga, P. C. and N. K. Chakravarti, "Optimal m-Stage Production Schedule," J. Can. Operations Res. Soc., Vol. **6**, 71–78 (1968).

[2] Dudek, R. A. and O. F. Teuton, Jr., "Development of M-Stage Decision Rule for Scheduling n Jobs Through M-Machines," Operations Research **12**, 471–497 (1964).

[3] Ignall, E. and L. Schrage, "Application of the Branch and Bound Technique to Some Flow-Shop Scheduling Problem," Operations Research, **13**, 400–412 (1965).

[4] Karush, W., "A Counterexample to a Proposed Algorithm for Optimal Sequencing of Jobs," Operations Research, **13**, 323–325 (1965).

[5] Smith, R. D. and R. A. Dudek, "A General Algorithm for Solution of the n-Job M-Machine Sequencing Problem of the Flow Shop, Operations Research, **15**, 71–82 (1967), and Errata, Operations Research, **17**, 756 (1969).

[6] Szwarc, W., "A Counterexample to a Solution Method of the m x n Job Sequencing Problem," Carnegie-Mellon University, School of Urban and Public Affairs (Feb. 1971).

THE FRACTIONAL FIXED-CHARGE PROBLEM

Y. Almogy and O. Levin

Technion—Israel Institute of Technology,
Haifa, Israel.

ABSTRACT

Fractional fixed-charge problems arise in numerous applications, where the measure of economic performance is the time rate of earnings or profit (equivalent to an interest rate on capital investment). This paper treats the fractional objective function, after suitable transformation, as a linear parametric fixed-charge problem. It is proved, with wider generality than in the case of Hirsch and Dantzig, that some optimal solution to the generalized linear fixed-charge problem is an extreme point of the polyhedral set defined by the constraints. Furthermore, it is shown that the optimum of the generalized fractional fixed-charge problem is also a vertex of this set. The proof utilizes a suitable penalty function yielding an upper bound on the optimal value of the objective function; this is particularly useful when considering combinations of independent transportation-type networks. Finally, it is shown that the solution of a fractional fixed-charge problem is obtainable through that of a certain linear fixed-charge one.

Introduction

We begin by describing a ship-routing problem which is typical of how fractional fixed-charge problems arise in practice. In a network of N ports, cargoes are available at every port for shipping within the network; their availability is assumed to be independent of time, i.e., of the frequency of calls at the port. Furthermore, it is assumed that there exists a "home port" at which all routes originate and terminate, and that cargoes are only destined to ports with a higher index, such that those ports which are visited are visited in ascending order of their indices. (Of these two assumptions, in practice, the former is always true, while the latter is not restrictive since it can be circumvented in most cases by introducing dummy ports into the network.) The route and loading plan for a ship (or a fleet of ships) have to be chosen so that profit per unit time is maximized. The net revenue is a separable function of the amounts of the (directed) cargoes carried by the ship on its route. Expenses consist of daily costs in port (which depend on the time spent there, in turn dependent on the amounts of loaded and unloaded cargoes), of daily costs at sea (which depend on the route chosen), and of fixed charges in port. The time required to cover a given route is a separable function of the amounts of the cargo carried, of the fixed times spent in port, and of sea-time on the route. The criterion for choosing a route is that the resulting profit per unit time (positive or negative) exceeds the fixed cost per unit time at the "home port."

The general constraints involved in this problem are those on cargo availability and on the cargo carrying capacity of the ship. These constraints are linear in the amounts of directed cargoes shipped, forming a close and bounded polyhedral set.

A route is defined in the following manner:

For every j, $j = 1, \ldots, N$, let

$$\delta_j = \begin{cases} 1 & \text{if the ship calls at port } j \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

From the above construction of the network, we have for the "home port" (designated as port 1 and $N+1$),

$$\delta_1 (= \delta_{N+1}) = 1.$$

Define the set

$$U(\delta) = U = \{i, j(i) | \delta_i = 1\}, \tag{2}$$

where for each $i (i \leq N)$ with $\delta_i = 1$,

$$j(i) = \min_{k>i} (k | \delta_k = 1). \tag{3}$$

Without going in detail into the significance of the various coefficients [12] the profit per unit time for a given route U has the general form

$$f = \frac{\alpha^T x - \delta^T \beta - \sum_{i,j \epsilon U} \hat{a}_{ij}}{\gamma^T x + \delta^T \xi + \sum_{i,j \epsilon U} a_{ij}}, \tag{4}$$

subject to the set of constraints

$$Ax \leq b;\ x \geq 0; \qquad \delta \text{ as defined in (1),}$$

where

x — a K-dimensional column vector of the amounts of cargoes taken from a port. Here K is the number of pairs (i, j) with $1 \leq i < j \leq N+1$. A typical component x_{ij} of x denotes the amount of cargo picked up at port i for delivery to port j. The constraints $Ax \leq b$ include representations both of cargo availability limits (e.g., upper bounds on the variables x_{ij}), and of the limited capacity of the ship.

α — a K-dimensional column vector of the net revenue (including time costs) per unit of directed cargo.

δ — an N-dimensional column vector, with components $\delta_j \epsilon \{0, 1\}$.

β — an N-dimensional column vector of the fixed time costs at each port.

γ — a K-dimensional column vector of the loading and unloading time per unit of directed cargo.

ξ — an N-dimensional column vector of the fixed times in port.

$\hat{a}_{ij}$ — the sea-time cost between ports i and j.

a_{ij} — the sea-time between ports i and j.

An analog, to the more general problem to be treated next, will now be identified. For each j, let I_j consist of all ordered pairs (i, j) with $i < j$ and all ordered pairs (j, i) with $j < i$. Then clearly

$$\sum_{(k,l) \epsilon I_j} x_{kl} = 0 \qquad \text{if} \qquad \delta_j = 0.$$

Without loss of generality we can require that ships not call at ports if the levels of cargo picked up there and cargo discharged there are both zero; then the "if" in the last display becomes "if and only if,"

$$\delta_j = \begin{cases} 1 & \text{if } \sum_{(k,l)\epsilon I_j} x_{kl} > 0 \\ 0 & \text{otherwise.} \end{cases}$$

For networks comprising two separate sets of ports, with the ports within each set closely spaced and the cargoes restricted to inter-set routes, the time at sea can be regarded as constant for all routes. Thus, setting

$$\sum_{i,j\epsilon U} \hat{a}_{ij} = \hat{a}; \qquad \sum_{i,j\epsilon U} a_{ij} = a,$$

problem (4) becomes a fractional fixed charge problem:

$$\text{Maximize } f = \frac{\alpha^T x - \delta^T \beta - \hat{a}}{\gamma^T x + \delta^T \xi + a} \tag{5}$$

subject to the set of constraints described earlier.

GENERAL FRACTIONAL FIXED-CHARGE PROBLEM

The general type of mathematical problem to be treated, illustrated by the ship-routing example above, involves

x, $\hat{c}$ and $\hat{d}$ — n-dimensional column vectors

δ, $\hat{t}$ and $\hat{T}$ — k-dimensional column vectors

$\hat{a}$ and $\hat{\epsilon}$ — given constants.

The general fractional fixed-charge problem is:

$$\text{Maximize } f(x) = \frac{\hat{c}^T x + \delta^T \hat{t} + \hat{a}}{\hat{d}^T x + \delta^T \hat{T} + \hat{\epsilon}}, \tag{6}$$

subject to

$$x\epsilon D; \qquad D = \{x \mid Ax \leqslant b; \qquad x \geqslant 0\}, \tag{7}$$

D being assumed bounded, and letting

$$\sum_{i\epsilon I_j} x_i \equiv x(I_j)$$

$$\delta_j = \begin{cases} 1 & \text{if } x(I_j) > 0 \qquad j = 1, \ldots, k \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

where the I_j are given subsets of the set of indices $N = \{1, \ldots, n\}$,

$$\bigcup_{j=1}^{k} I_j \subseteq N. \tag{9}$$

For the problem to be meaningful, it is required that

$$\hat{d}^T x + \delta^T \hat{T} + \hat{\epsilon} > 0 \qquad \text{for every } x\epsilon D, \tag{10}$$

furthermore, we impose additional requirements

$$\hat{\epsilon} > 0; \qquad \hat{d} \geqslant 0, \tag{11}$$

expressing that the time measure derived from any physical or economic measure cannot be negative (for well-defined variables). Subject to these assumptions, the objective function (6) can be reformulated:

$$f(x)=\frac{c^Tx+\delta^Tt+a}{d^Tx+\ \delta^TT+1} \tag{12}$$

or, in different form,

$$f(x)=[c^T-f(x)d^T]x+\delta^T[t-f(x)T]+a. \tag{13}$$

Hereafter, we regard the value of $f(x)$ as a parameter f, which can be varied irrespective of whether it is attainable on D.

For a fixed value of the parameter f, the problem:

$$\text{Maximize } \{[c^T-fd^T]x+\delta^T[t-fT]+a=P(f)^Tx+\delta^Tq(f)+a\overset{d}{=}F(x,f)=F(x)\} \tag{14}$$

subject to $x\epsilon D$ and δ satisfying (8), is a linear generalized fixed-charge problem, to be referred to as problem (F).

To ensure that the maximum for problem (F) exists (i.e., that there is not an unattained supremum), it is assumed that

$$q_j(f)\leq 0 \qquad j=1,\ \ldots,\ k, \tag{15}$$

($q_j(f)>0$ would mean in general a "bonus" given for an infinitesimal positive level of the activities in I_j).

Consider now a different problem:

$$\text{Maximize } \{P(f)^Tx+\delta^Tq(f)+a-\sum_{j=1}^{k}M_jx(I_j)(1-\delta_j)\overset{d}{=}G(x,\delta,M)\} \tag{16}$$

subject to $x\epsilon D$; $\delta\epsilon\Delta$, where

$$\Delta=\{\delta|0\leq\delta_j\leq 1; \qquad j=1,\ \ldots,\ k\} \tag{17}$$

M — some nonzero, finite, k-dimensional vector.

This problem is referred to as problem (G). For a given vector M, this is a program with quadratic (in x and δ) nondefinite objective function subject to linear constraints.

For problems (F) and (G), the following holds: For each $x\epsilon D$ and the associated $\delta(x)$ of (8), there exists an M, such that

$$F(x)=G(x,\delta(x),M)=\max_{\delta\epsilon\Delta|x}G(x,\delta,M). \tag{18}$$

PROOF: Without loss of generality, assume

$$\begin{aligned} a&=0\\ x(I_j)&=0 \qquad j=1,\ \ldots,\ l\\ x(I_j)&>0 \qquad j=l+1,\ \ldots,\ k\\ x_i&=0 \qquad i=1,\ \ldots,\ p\\ x_i&>0 \qquad i=p+1,\ \ldots,\ n. \end{aligned}$$

Simplifying the notation, we have

$$F(x)=\sum_{i=p+1}^{n} p_i x_i+\sum_{j=l+1}^{k} q_j.$$

We denote

$$\min_{l+1\le j\le k} x(I_j)=\epsilon>0 \tag{19}$$

and choose a nonnegative vector M, such that,

$$M_j\epsilon>|q_j| \qquad j=l+1,\ \ldots,\ k. \tag{20}$$

For a given vector $x\epsilon D$ and a vector M chosen as above, the function $G(x, \delta, M)$ is linear in δ, defined on a bounded polyhedral set Δ; thus, its maximum is attained at an extreme point of Δ.

(I) $x(I_j)=0$ leads to $\delta_j=0$, because the coefficients of the variables δ_j are nonpositive.

(II) $x(I_j)>0$ leads to $\delta_j=1$, as a result of introducing a penalty function with the appropriate vector M.

From the above, it follows that

$$\max_{\delta\epsilon\Delta|x} G(x, \delta, M)=\sum_{i=p+1}^{n} p_i x_i+\sum_{j=l+1}^{k} q_j=F(x),$$

proving what was stated.

In particular

$$\max_{x\epsilon D} F(x)=F(x^*)=\max_{\delta\epsilon\Delta|x^*} G(x^*, \delta, M^*)=G(x^*, \delta^*, M^*).$$

Having proved this property, we state theorem 1:

THEOREM 1: Some optimal solution to problem (F) is at an extreme point of D.

PROOF: Suppose that x^*, an optimal solution to problem (F), is not an extreme point of D.

Without loss of generality, assume:

$$\begin{aligned} a&=0 \\ x^*(I_j)&=0 \qquad j=1,\ \ldots,\ l \\ x^*(I_j)&>0 \qquad j=l+1,\ \ldots,\ k. \end{aligned}$$

Then

$$\begin{aligned} \delta_j^*&=0 \qquad j=1,\ \ldots,\ l \\ \delta_j^*&=1 \qquad j=l+1,\ \ldots,\ k, \end{aligned}$$

and the vector M^* is chosen, such that,

$$\begin{aligned} M_j^* \min_{x\epsilon B}\{\min_{j|x(I_j)>0} x(I_j)\}&>|q_j| \qquad j=1,\ \ldots,\ l \\ M_j^* \min_{l+1\le j\le k} x^*(I_j)&>|q_j| \qquad j=l+1,\ \ldots,\ k, \end{aligned}$$

where B is the set of all extreme points of D.

It is clear that for these δ^* and M^*

$$F(x^*)=G(x^*,\delta^*,M^*)\leq \max_{x\in D|\delta^*} G(x,\delta^*,M^*)=G(\hat{x},\delta^*,M^*)$$

for some extreme point $\hat{x}$ of D.

Without loss of generality assume:

$$\begin{aligned}
&\hat{x}(I_j)=0 \qquad j=1,\ \ldots,\ m\\
&\hat{x}(I_j)>0 \qquad j=m+1,\ \ldots,\ l\\
&\hat{x}(I_j)=0 \qquad j=l+1,\ \ldots,\ p\\
&\hat{x}(I_j)>0 \qquad j=p+1,\ \ldots,\ k.
\end{aligned}$$

It follows that

$$M_j^*\hat{x}(I_j) > |q_j| \qquad j=m+1,\ \ldots,\ l,$$

but it does *not* necessarily follow that

$$M_j^* \min_{p+1\leq j\leq k} \hat{x}(I_j) > |q_j| \qquad j=p+1,\ \ldots,\ k,$$

so the lemma's proof does not apply to $\hat{x}$ and M^*. A vector $\hat{M}$ is chosen, such that

$$\hat{M}_j=M_j^* \qquad j=1,\ \ldots,\ p$$

$$\hat{M}_j \min_{p+1\leq j\leq k} \hat{x}(I_j) > |q_j| \qquad j=p+1,\ \ldots,\ k,$$

so that

$$M_j^*\hat{x}(I_j)(1-\delta_j^*)=\hat{M}_j\hat{x}(I_j)\ (1-\delta_j^*)$$

for $j=p+1,\ \ldots,\ k$ because both sides are zero, and for $j=1,\ 2,\ \ldots,\ p$ because $\hat{M}_j=M_j^*$. Thus, since the lemma's proof applies to $\hat{x}$ and $\hat{M}$,

$$G(\hat{x},\ \delta^*,\ M^*)=G(\hat{x},\ \delta^*,\ \hat{M}\ )\leq \max_{\delta\in\Delta|\hat{x}} G(\hat{x},\ \delta,\ \hat{M})=G(\hat{x},\ \hat{\delta},\ \hat{M})=F(\hat{x}).$$

Combining both inequalities, we get

$$(21)\quad F(x^*)=G(x^*,\delta^*,M^*)\leq \max_{x\in D|\delta^*} G(x,\delta^*,M^*)=G(\hat{x},\delta^*,M^*)\leq \max_{\delta\in\Delta|\hat{x}} G(\hat{x},\delta,\hat{M})=G(\hat{x},\hat{\delta},\hat{M})=F(\hat{x})$$

so that $\hat{x}$, and extreme point of D, is an optimal solution to problem (F).

A different proof of the theorem was given, for a less general case, by Hirsch and Dantzig [8].

For problems having a unimodular coefficient matrix (transportation, etc.) $\epsilon \geq 1$, and the vector M^* is readily obtainable

$$(22)\qquad M_j^* > |q_j| \qquad j=1,\ \ldots,\ k.$$

From the above, the following can now be stated:

THEOREM 2: Some optimal solution to the general linear fractional fixed-charge problem (6) is an extreme point of D.

PROOF: Let x^* be an optimal solution to problem (6). Let $\hat{x}$ be an optimal solution to problem (F) with parameter-level $f=f^*=f(x^*)$; by Theorem 1 we may assume that $\hat{x}$ is an extreme point of D.

It follows that

$$f^*=F(x^*,f^*)\leq F(\hat{x},f^*)=P(f^*)^T\hat{x}+\delta^T q(f^*)+a, \tag{23}$$

yielding $f(x^*)=f^*\leq f(\hat{x})$. Thus $\hat{x}$, an extreme point of D, is an optimal solution to problem (6).

BOUNDS ON OPTIMAL SOLUTION

Slightly modifying the procedure used above, define $\epsilon>0$ by

$$\epsilon=\min_{x\in B}\{\min_{j|x(I_j)>0} x(I_j)\}. \tag{24}$$

Let $\hat{M}$ satisfy

$$\hat{M}_j\epsilon>|q_j| \qquad j=1, 2, \ldots, k. \tag{25}$$

Then we know that, for any $x^*\epsilon B$,

$$F(x^*)=\max_{\delta\epsilon\Delta} G(x^*, \delta, \hat{M}). \tag{26}$$

Since $G(x^*, \delta, \hat{M})$ is linear in δ, it follows that

$$F(x^*)\leq \max_{\delta\epsilon E(\Delta)} G(x^*, \delta, \hat{M}), \tag{27}$$

where $E(\Delta)$ denotes the set of extreme points of Δ.

It follows that

$$F(x^*)\leq \max_{\delta\epsilon E(\Delta)} \max_{x\epsilon D} F(x, \delta, \hat{M}). \tag{28}$$

Since some optimal solution of problem (F) is an $x^*\epsilon B$, this gives an upper bound on the optimal value for the linear fixed-charge problem. For each $\delta\epsilon E(\Delta)$, $\max_{x\epsilon D} G(x, \delta, \hat{M})$ can be found by solving the associated linear program. The practical difficulty here lies in the number, 2^k, of members of $E(\Delta)$. We can settle for a cruder, but more accessible bound

$$F(x^*)\leq \max_{\delta\epsilon E(\Delta)} \max_{x\epsilon D} G(x, \delta, \hat{M}), \tag{29}$$

where Δ' is a convex set containing Δ whose extreme points are readily identifiable and less numerous than those of Δ. In particular, we can take

$$\Delta'=\{\delta \mid \sum_{j=1}^{k}\delta_j\leq k; \qquad \text{all } \delta_j\geq 0\}, \tag{30}$$

whose extreme points are the zero vector and the vectors $\{\delta_j^k\}_{j=1}^k$, where δ_j^k has all components zero except that its jth component is k.

The terms $\max_{x\epsilon D} G(x, \delta_j^k, \hat{M})$ are readily obtainable solutions of linear programming problems. The upper bound (29) is particularly useful where a combination of two or more fixed-charge networks is considered, as in numerous practical transportation or shipping problems. For a combination of L networks, the lth consisting of k_l fixed-charge sets, the upper bound on the optimal solution becomes

$$F(x^*)\leq \max_{j} \max_{x\epsilon D} G(x, \delta_j^s, \hat{M}) \tag{31}$$

$$s=\sum_{l=1}^{L} k_l.$$

Bounds of a different nature can be found for the fractional fixed-charge problem (13). The function

$$Z(f)=\max_{x\in D}\{[c^T-fd^T]x+g(f)\}, \tag{32}$$

where

$$g(f)=\sum_{j=1}^{k}[t_j-fT_j]+a$$

is convex and monotone nonincreasing in f everywhere, and monotone-decreasing in f in some range [2]

$$[-\infty\leqslant f\leqslant \bar{f}<\infty].$$

The linear fractional programming problem associated with (32) is solved by finding a value f^* such that

$$f^*=Z(f^*). \tag{33}$$

This is a problem of finding a root of a convex monotone-decreasing function. Once f^* is found, the value of the parameter f can be varied and $Z(f)$ determined in the range

$$[f^*,\bar{f}],$$

using the sensitivity analysis feature available in most L.P. computer codes. By utilizing this feature, upper bounds are readily obtained for the optimal solutions to fractional problems with $m\leqslant k$ fixed-charge sets. These bounds are the values of the parameter f for which the following holds:

$$f=\max_{x\in D}[c^T-fd^T]x+\sum_{j\in P}[t_j-fT_j]+a \tag{34}$$

where the set P consists of any given combination of m fixed-charge sets I_j.

METHODS OF SOLUTION

Apart from approximate methods [6], [7], [9], there are two principal approaches to linear fixed-charge problems. One is the mixed-integer programming approach [4], which is not efficient in the fractional case, as each improvement in the value of the objective function involves a new parameter in the linear functional and calls for a new program. The other consists in ranking the extreme points of D and finding at each step a lower bound on the optimal value of the (linear) objective function [11]. A method such as that of Pollatschek and Avi-Itzhak [12], which uses the linear functional as cutting plane, consists in checking at each extreme point $\hat{x}$ whether

$$f^*\leqslant[c^T-f^*d^T]\hat{x}+\delta^T[t-f^*T]+a.$$

It is thus clear that the linear fractional fixed-charge problem is of the same degree of difficulty as the linear one. Furthermore, solving it is equivalent to solving a linear fixed-charge problem, namely (14).

An interesting (although not particularly suitable in the fractional case) method consists in representing the (linear) fixed-charge problem as a linear program with logical constraints. The problem becomes

$$\text{Maximize } \{P^Tx+\delta^Tq\}$$

subject to

$$x \in D$$

and

$$(x(I_j) = 0)\ \gamma\ (\delta_j = 1) \qquad j = 1, \ldots, k.$$

If the number of fixed-charge sets is small compared with that of variables (as in the case in the shipping problem [2]), and a relatively good lower bound is available for the optimal value of the objective function, a "relaxation-branching" procedure suggested by Rado [13], utilizing the dual-simplex procedure (which is a relaxation procedure), may prove to be an efficient algorithm.

It seems that for this problem, ranking the extreme points of D is, at this stage, the most efficient method of solution. In this case, also, the linear and the fractional problems are of the same degree of difficulty.

BIBLIOGRAPHY

[1] Almogy, Y. and Levin, O., "A Class of Fractional Programming Problems," Accepted for publication in Opns. Res.

[2] Almogy, Y. and Levin, O., "Parametric Analysis of a Multi-Stage Stochastic Shipping Problem," Proceedings, 5th IFORS Conference, Venice, 1969.

[3] Balinski, M. L., "Fixed Cost Transportation Problems," Nav. Res. Log. Quart. **8**, 41–54 (1961).

[4] Balinski, M. L., "Integer Programming Methods, Uses, Computation," Management Science **12**, 253–313 (1965).

[5] Charnes, A. and W. W. Cooper, "Programming with Linear Fractional Functionals," Nav. Res. Log. Quart. **9**, 181–196 (1962).

[6] Cooper, L. and C. Drebes, "An Approximate Solution Method for the Fixed Charge Problem," Nav. Res. Log. Quart. **14**, 101–113 (1967).

[7] Denzler, D. R., "An Approximate Algorithm for the Fixed Charge Problem," Nav. Res. Log. Quart. **16**, 411–416 (1969).

[8] Hirsch, W. M. and G. B. Dantzig, "The Fixed Charge Problem," Nav. Res. Log. Quart. **9**, 413–424, (1968).

[9] Kuhn, H. W. and W. J. Baumol, "An Approximate Algorithm for the Fixed-Charge Transportation Problem," Nav. Res. Log. Quart. **9**, 1–16 (1962).

[10] Martos, B., "Hyperbolic Programming," Nav. Res. Log. Quart. **11**, 135–155 (1964).

[11] Murty, K. G., "Solving the Fixed Charge Problem by Ranking the Extreme Points," Operations Research **16**, 268–279 (1968).

[12] Pollatschek, M. A. and B. Avi-Itzhak, "Sorting Feasible Basic Solutions of a Linear Program," paper presented at joint ORSIS–ORSA Conference, Tel-Aviv, 1969.

[13] Rado, F., "Une Algorithme pour résoudre certains problemes de programation mathématique," Mathematica **6**, 105–116 (1964).

A HYBRID ALGORITHM FOR THE ONE MACHINE SEQUENCING PROBLEM TO MINIMIZE TOTAL TARDINESS*

V. Srinivasan

The University of Rochester
Rochester, New York

ABSTRACT

In a recent paper, Hamilton Emmons has established theorems relating to the order in which pairs of jobs are to be processed in an optimal schedule to minimize the total tardiness of performing n jobs on one machine. Using these theorems, the algorithm of this paper determines the precedence relationships among pairs of johs (whenever possible) and eliminates the first and the last few jobs in an optimal sequence. The remaining jobs are then ordered by incorporating the precedence relationships in a dynamic programming framework. Propositions are proved which considerably reduce the total computation involved in the dynamic programming phase. Computational results indicate that the solution time goes up less than linearly with the size (n) of the problem. The median solution time for solving 50 job problems was 0.36 second on UNIVAC 1108 computer.

I. INTRODUCTION

The one-machine sequencing problem to minimize total tardiness may be stated as follows:

We are given a set of jobs $\{1, 2, \ldots, n\}$ whose sequence-independent processing times $p_1, p_2, \ldots, p_n$ and due dates $d_1, d_2, \ldots, d_n$ are known. The jobs are available simultaneously at time zero and are to be sequenced on one machine. The set-up times are independent of the sequence and consequently are assumed to be included in the processing times.

Let $\{(1), (2), \ldots, (n)\}$ be a permutation of the integers $\{1, 2, \ldots, n\}$ denoting a particular sequence of the n jobs. (Thus $(2)=4$ if the fourth job is scheduled second.) The completion time of the (i)th job $c_{(i)}=p_{(1)}+p_{(2)}+\ldots+p_{(i)}$ and its lateness $l_{(i)}=c_{(i)}-d_{(i)}$. Minimizing total lateness may not be a good objective since it gives equal credit for finishing a job early as it penalizes for finishing a job late. Tardiness considers only positive lateness (jobs which are not completed by their due dates) i.e., $t_{(i)} = \max [l_{(i)}, 0]$ so that minimizing total tardiness may be a more desirable objective in many situations. Thus for a sequence $\{(1), (2), \ldots, (n)\}$ the total tardiness T is given by

$$(1) \qquad T=\max [0, (p_{(1)}-d^z{}_{(1)})]+\max [0, (p_{(1)}+p_{(2)}-d_{(2)})]+\cdots+\max [0, (p_{(1)}+p_{(2)}+\cdots+p_{(n)}-d_{(n)})]$$

The objective is to find a sequence that minimizes T. Other reasons for using T rather than other measures of performance are discussed in [2].

II. PREVIOUS RESEARCH

For the general case, where no restrictive assumptions are imposed on the processing times and due dates, several algorithms have been proposed for minimizing total tardiness [4–6, 8–10]. The algorithm of Schild and Fredman [9] cannot guarantee optimality [3] and no computational results have

*This report was prepared when the author was a member of the Management Sciences Research Group, Carnegie-Mellon University. Reproduction in whole or in part is permitted for any purpose of the U.S. Government.

been published about the quality of the solution. The network algorithm due to Elmaghraby [4], which a significant improvement over complete enumeration, is laborious and applicable only to smaller problems. A dynamic programming formulation for the general case of nonlinear loss functions has been given by Held and Karp [6], Lawler [8], and Schild and Fredman [10]. Computer memory requirements restrict the size of the problem that can be handled by this formulation, for example, the maximum problem that could be solved on the IBM 7090 is 13 jobs without resorting to any bulk storage [6].

In Ref. [5] Emmons has proved theorems that establish the relative order in which pairs of jobs are processed in an optimal schedule. If for some pair(s) of jobs the conditions of the theorems are not satisfied, then one obtains only a partial order of the jobs in the optimal sequence. If, for instance, nothing is known about the relative order of the jobs (i, j) then Emmons suggests branching into two subproblems by assuming i precedes j and vice versa. For a large problem, the size of tree so obtained can quickly get out of hand. Since no computational results are available for Emmons' algorithm it could not be compared with the present one.

III. THE HYBRID ALGORITHM

The algorithm proposed in this paper is hybrid in the sense that it incorporates theoretical results into the dynamic programming framework to reduce the total computational effort.

Outline of the Algorithm:

PHASE 1: Jobs which have excessively long due dates (to be defined shortly) are directly identified as the last jobs in an optimal sequence.

PHASE 2: Theorems 1 and 2 of Emmons [5] are repeatedly used to obtain as many precedence relationships among the jobs as possible. In general, the first and the last few jobs of the optimal sequence can be directly identified in this process.

PHASE 3: The remaining jobs are ordered by incorporating the precedence relationships in a dynamic programming framework [6, 8, 10] modified to take into account the precedence relationships determined in phase 2.

Detailed Description of Phases 1, 2, and 3:

PHASE 1: If $d_i > P = \sum_{j=1}^{n} p_j$, then there exists an optimal schedule in which i is processed last [5, p. 703].

The result is obvious since by the property that $d_i > P$, job i is never tardy in any sequence and hence can safely be sequenced last. In doing so, P is reduced to $P - p_i$ and another job may become eligible for removal as the last job among the remaining $n-1$ jobs. After repeating this procedure as many times as possible, we update n to be the remaining number of jobs.*

PHASE 2: Hereafter we will assume that the n remaining jobs are arranged in the order of non-decreasing processing times ($p_1 \leq p_2, \ldots, \leq p_n$) and in the case of equality, in the order of non-decreasing due dates. Thus $i < j$ implies $p_i < p_j$ or $p_i = p_j$ and $d_i \leq d_j$. Let N be the set of jobs $\{1, 2, \ldots, n\}$ arranged in that order.

---

*If at any stage, there are k jobs ($k > 1$) which satisfy this elimination criterion, then one can arbitrarily choose any one of them to be the last. The remaining $(k-1)$ of these jobs will get eliminated in successive stages. The order is not important since each of these jobs should have zero tardiness in the optimal sequence.

We use the notation $i \leftarrow j$ which may be read as "i precedes j" to mean that there exists an optimal schedule in which i precedes j. Let A_i and B_i be the sets of indices of all jobs that at any point in the algorithm have been shown to come after and before i, respectively. We will denote the number of jobs in a set S as $|S|$.

To keep track of the precedence relationships, we introduce the notion of a *Precedence Matrix M* (of size $n \times n$) where $m_{ij}=1$ if $i \leftarrow j$ and $m_{ij}=0$ otherwise. The diagonal elements m_{ii} of this matrix do not have any significance and are marked '—'. By definition, $A_i=\{j|m_{ij}=1\}$ and $B_i=\{j|m_{ji}=1\}$. Thus $|A_i|$ will be given by the number of 1's in the ith row and $|B_i|$ will be given by the number of 1's in ith column of the matrix M.

For convenience of the reader Emmons' Theorems 1 and 2 are restated below using the notation of the precedence matrix. Computational results revealed that the additional precedence relationships obtained by applying Emmons' Theorem 3 were negligible and hence this theorem was not used.

THEOREM 1: For any two jobs j and k with $j < k$, if $d_j \leq \max\left(p_k + \sum_{i|m_{ik}=1} p_i, d_k\right)$, then $m_{jk} = 1$.

THEOREM 2: For any two jobs j and k with $j<k$, if $d_j > \max\left(p_k + \sum_{i|m_{ik}=1} p_i, d_k\right)$ and $d_j + p_j \geq \sum_{i|m_{ki}=0} p_i$, then $m_{kj}=1$.

We first apply Theorems 1 and 2 to all pairs (j, k) with $j < k$. If any precedence relationships are obtained, then reapplication of the theorems may generate additional precedence relationships. This iterative process terminates when no new precedence relationships are obtained in a particular iteration.

Any theorems regarding precedence relationships that may subsequently become known could be readily built into this framework provided they result in net computational gain.

It is obvious that if $|A_i|=n-1$ then i is clearly the first job in the optimal sequence. Thus i can be removed from the set N and n redefined as $n-1$. The due dates of the jobs in the set N become d_j-p_i to account for the fact that i has already been scheduled. Similarly, if $|B_i|=n-1$, i is sequenced last, and N and n are accordingly redefined. Whenever a job is identified as either first or last, additional precedence relationships may be generated by reapplying Theorems 1 and 2.

PHASE 3: We first describe the well known dynamic programming approach to this problem [6, 8, 10] and then state some propositions which will considerably reduce the computation involved.

Let J denote any set of k jobs scheduled last. Let $s(J)$ denote the "Earliest Start Time" of the jobs in J i.e., $s(J)=\sum_{i\in\tilde{J}} p_i$, where $\tilde{J}=N-J$ denotes the first $(n-k)$ jobs. Bellman's Principle of Optimality [1] may now be stated: If the schedule is optimal, it has the property that regardless of the order in which the first $n-k$ jobs are performed, the remainder of the schedule constitutes an optimal schedule of the subset of jobs J subject to the restriction that none of the jobs are commenced before $s(J)$.

Let $T(J)$ denote the minimum total tardiness of performing the jobs in J subject to the restriction that no job in J is commenced before $s(J)$.

To obtain the recursive relation of the dynamic programming formulation we reason as follows: $T(J|i)$, the minimum tardiness of J given that $\{i\}$ is sequenced first among the jobs in J, is:

$$T(J|i)=\max(0, p_i+s(J)-d_i)+T(J-\{i\}). \tag{2}$$

Here the first term denotes the contribution to total tardiness of job $\{i\}$ and the second term gives the minimum tardiness of the remaining jobs in J. Removing the condition on $\{i\}$, we get

$$T(J) = \min_{i \in J} [\max (0, p_i + s(J) - d_i) + T(J - \{i\})], \tag{3}$$

where $T(\phi) = 0$ for the empty set ϕ.

The recurrence relation (3) is to be repeated for $|J| = 1, 2, \ldots, n$ to obtain the optimal solution. At the kth stage (i.e., $|J| = k$) one would normally evaluate $T(J)$ for all the $\binom{n}{k}$ combinations of jobs. We now prove three propositions which considerably reduce this number of evaluations.

PROPOSITION 1: Let $R = \{i: |A_i| \geq k\}$. Let $N_1 = N - R$. Then for the kth stage optimization one need consider only the n_1 jobs in N_1.

PROOF: $|A_i| \geq k$, by definition, means that there exists an optimal schedule in which i precedes at least k jobs. Thus i need not be considered for the kth stage optimization.

PROPOSITION 2: Let $Q = \{i: |A_i| = k - 1\}$. Let $N_2 = N_1 - Q$ contain n_2 jobs. Then in the set of J's in which i appears, one need consider only that $J = \{i\} \cup A_i$ for the kth stage. Furthermore

$$T(J) = \max (0, p_i + s(J) - d_i) + T(A_i). \tag{4}$$

PROOF: $|A_i| = k - 1$ implies that there exists an optimum schedule in which i precedes the $k - 1$ jobs in A_i. Thus $\{i\}$ is the first job in the optimal schedule for J so that $T(J) = T(J|i)$. Eq. (4) directly follows from (2) since $J - \{i\} = A_i$.

As a result of the above two propositions the total number of combinations for the kth stage are reduced from $\binom{n}{k}$ to $\binom{n_2}{k} + (n_1 - n_2)$. Proposition 3 below further reduces this number.

PROPOSITION 3: Let i be a job, such that $m_{ij} = 1$ for $j = i_1, i_2, \ldots, i_l; l < k - 1$. Then in the set of J's in which i appears, one need choose only those in which i appears together with jobs $i_1, i_2, \ldots, i_l$.

PROOF: The proof of this proposition is similar to that of Proposition 2.

We now consider the functional equation, Eq. (3), where for a given J, one would normally have to evaluate all the k alternatives for the first job in J. Proposition 4 below reduces this number of evaluations.

PROPOSITION 4: Let $L = \{i: |B_i| \geq n - k + 1\}$. Then only the jobs in $N_3 = N - L$ need be considered as candidates for the first job of the kth stage optimization for J.

PROOF: $|B_i| \geq n - k + 1$ means that there is an optimal schedule in which at least $n - k + 1$ jobs precede i; i.e., i is one of the last $(k - 1)$ jobs in the optimal schedule. Hence i need not be considered for the first position in a k job problem. Thus all potential candidates for the first job are in $N_3 = N - L$. Using this result we may rewrite (3) as

$$T(J) = \operatorname*{Min}_{i \in N_3 \cap J} [\max (0, p_i + s(J) - d_i) + T(J - \{i\})]. \tag{5}$$

IV. AN EXAMPLE

We illustrate the application of the hybrid algorithm on a 10-job problem for which the necessary data is provided in Table 1 below:

TABLE 1. *Data For a 10-Job Problem*

Jobs	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
p	7	1	2	3	9	8	4	3	6	9
d	4	13	10	10	5	12	11	27	6	60

PHASE 1: $P=\sum_{1}^{10} p_j=52$. $d_J=60>P$. Hence J is the last job. For the remaining nine jobs $P=52-9=43$. There is no other job which has $d_i>P$. We proceed to Phase 2.

PHASE 2: We arrange the remaining nine jobs in the order $i<j$ to get the data of Table 2. Note that job J is no longer a part of the list.

TABLE 2. *Jobs Rearranged in the Order* j < k

Job No.	1	2	3	4	5	6	7	8	9
Job name	*B*	*C*	*D*	*H*	*G*	*I*	*A*	*F*	*E*
p	1	2	3	3	4	6	7	8	9
d	13	10	10	27	11	6	4	12	5

Upon repeated application of Theorems 1 and 2 given earlier to all pairs of jobs, we obtain the precedence matrix shown in Figure 1. It is clear that E is to be scheduled last among the remaining nine jobs.* On removing E from the list (i.e., crossing out both the row and column of the precedence matrix, corresponding to job E), F gets identified as the last among the remaining eight jobs ($|B_F|=7$).

Job name		B	C	D	H	G	I	A	F	E
	Job No.	1	2	3	4	5	6	7	8	9
B	1	.....			1			1	1	1
C	2	1	.....	1	1	1		1	1	1
D	3	1		.....	1	1		1	1	1
H	4				.....				1	1
G	5				1	.....		1	1	1
I	6				1		.....	1	1	1
A	7				1			.....	1	1
F	8								.....	1
E	9									.....

FIGURE 1. Precedence matrix for the nine jobs after repeated application of Theorems Nos. 1 and 2

*Refer to the description of Phase 2. Here $|B_E|=8$ for the nine-job sub problem.

Similarly H and A get eliminated on successive steps and no further jobs can be directly identified as first or last. Thus for the 10-job problem the last 5 jobs in the optimal sequence are $\{A, H, F, E, J\}$ in that order. The remaining five jobs $\{B, C, D, G, I\}$ and the precedence relationships that pertain to them are displayed in the precedence matrix of Figure 2. These jobs are renumbered as {1, 2, 3, 4, 5}.

Job name		B	C	D	G	I
	Job No.	1	2	3	4	5
B	1	.....				
C	2	1	.....	1	1	
D	3	1		.....	1	
G	4				.....	
I	5					.....
A_i		0	3	2	0	0
B_i		2	0	1	2	0
p_i		1	2	3	4	6
d_i		13	10	10	11	6

FIGURE 2. Precedence matrix for the five jobs in the dynamic programming phase

PHASE 3: We now find the optimal sequence of the remaining five jobs. In each of the five stages of the dynamic program, we apply Propositions 1 through 4 to reduce the computational burden. In this particular example, Proposition 3 does not turn out to be useful though such is not the case for larger problems.

STAGE 1:

$$(k=1): N=\{1,2,3,4,5\}; R=\{2,3\} \text{ (Proposition 1);}$$

Thus $N_1=N-R=\{1, 4, 5\}=Q$ (Proposition 2). We illustrate the calculations for the particular case when $J=\{1\}$. $s(J)=p_2+p_3+p_4+p_5=2+3+4+6=15$. Applying (4), we get $T(J)=\max(0, 1+15-13)+T(\phi)=\max(0, 3)+0=3$. Similar calculations are repeated for $J=\{4\}$ and $J=\{5\}$ to get the results shown in Table 3. Since for this stage $|J|=1$, the optimal order of the jobs in J is J itself.

TABLE 3. $k=1$

J	Optimal order	$T(J)$
{1}	(1)	3
{4}	(4)	5
{5}	(5)	10

TABLE 4. $k=2$

J	Optimal order	$T(J)$
{1, 4}	(1, 4)	5
{1, 5}	(1, 5)	10
{4, 5}	(4, 5)	10

STAGE 2:

$$R=\{2, 3\} \qquad N_1=\{1, 4, 5\}$$

$$Q=\phi \qquad N_2=N_1-Q=\{1, 4, 5\}$$

$$L=\phi \qquad N_3=N-L=\{1, 2, 3, 4, 5\}$$

We have to consider all pairs of jobs in the set N_2 viz., $\{1, 4\}$, $\{1, 5\}$, and $\{4, 5\}$ for J. We illustrate the calculations below for $J=\{1, 4\}$: $s(J)=p_2+p_3+p_5=2+3+6=11$. We now use (5) to get $T(J)=\text{Min}$ $[\{\max(0, 1+11-13)+T(4)\}, \{\max(0, 4+11-11)+T(1)\}]$. Here $N_3 \cap J=\{1, 4\}$ so that the first and second alternatives in the above expression correspond to using $\{1\}$ and $\{4\}$, respectively, as first job in J. $T(J)=\text{Min}\ [5, 7]=5$. Thus optimal order for J is (1, 4). The results of Stage 2 are shown in Table 4.

STAGE 3:

$$R=\{2\}; \qquad N_1=\{1, 3, 4, 5\}; \qquad Q=\{3\}; \qquad N_2=\{1, 4, 5\}$$

$$L=\phi; \qquad N_3=\{1, 2, 3, 4, 5\}.$$

Corresponding to the job in Q viz., $\{3\}$, the two jobs that it precedes are $\{1\}$ and $\{4\}$. From Table 4, $\{1, 4\}$ is an optimal schedule. Thus the optimal order for the set of jobs $\{1, 3, 4\}$ is (3, 1, 4); $s(J)=p_2+p_5=8$. From (4) we get $T(J)=\max[0, 3+8-10]+T\{1, 4\}=1+5=6$.

From the set N_2, only one three-job combination can be obtained; $J=\{1, 4, 5\}$. $s(J)=p_2+p_3=5$. In applying (5) all the three candidates in J viz., 1, 4, 5 could be first jobs since they all belong to N_3.

$$T(J)=\text{Min}[\{\max(0, 1+5-13)+10\}, \{\max(0, 4+5-11)+10\}, \{\max(0, 6+5-6)+5\}]$$

$$=\text{Min}[10, 10, 10]=10.$$

Here the first, second, and third alternatives correspond to placing $\{1\}$, $\{4\}$, and $\{5\}$ as the first jobs. Since all three alternatives yield the same solution, we can arbitrarily choose one of them — say (1, 4, 5). The results of Stage 3 are shown in Table 5.

TABLE 5. k = *3*

J	Optimal order	*T(J)*
{1, 3, 4}	(3, 1, 4)	6
{1, 4, 5}	(1, 4, 5)	10

TABLE 6. k = *4*

J	Optimal order	*T(J)*
{1, 2, 3, 4}	(2, 3, 1, 4)	6
{1, 3, 4, 5}	(5, 3, 1, 4)	8

STAGE 4:

$$R=\phi; \qquad N_1=\{1, 2, 3, 4, 5\}; \qquad Q=(2); \qquad N_2=\{1, 3, 4, 5\};$$

$$L=\{1, 4\}; \qquad N_3=\{2, 3, 5\}.$$

Corresponding to $\{2\}$ in Q, the optimal schedule $J=\{2, 1, 3, 4\}$ is clearly (2, 3, 1, 4) and by using (4) we get $T(J)=6$. From N_2 we can form only one four-job combination; $J=\{1, 3, 4, 5\}$. By (5) then

$$(1, 3, 4, 5)=\text{Min}\ [\{\max(0, 3+2-10)+10\}, \{\max(0, 6+2-6)+6\}]=8.$$

Here the first and second alternatives correspond to jobs $\{3\}$ and $\{5\}$ in $N_3 \cap J$. The results for Stage 4 are shown in Table 6.

STAGE 5:

$$R=\phi; \qquad Q=\phi; \qquad N_1=N_2=\{1, 2, 3, 4, 5\};$$

$$L=\{1, 3, 4\}; \qquad N_3=\{2, 5\}; \qquad J=N_2=\{1, 2, 3, 4, 5\}; \qquad s(J)=0.$$

$J \cap N_3=\{2,5\}$ so that by applying (5) we obtain

$$T(J)=\text{Min}\,[\{\max(0,\ 2-10)+8\},\ \{\max(0,\ 6-6)+6\}]=6.$$

Thus the optimal sequence for the five-job problem $=\{5, 2, 3, 1, 4\}=\{I\ C\ D\ B\ G\}$. Thus the optimal sequence for the 10-job problem is $\{I\ C\ D\ B\ G\ A\ H\ F\ E\ J\}$ with a total tardiness of 85.

V. COMPUTATIONAL RESULTS

The algorithm of Section III was coded* in Fortran V and tested with a number of randomly generated problems using the UNIVAC 1108. The results are very promising; however, at this point there are no computational results for the other algorithms [4, 5, 6, 8, 9, 10] available for comparison.

The processing times and the due dates were drawn from a bivariate normal distribution confined to the positive quadrant. The distribution is completely specified by mean processing time and due date μ_p and μ_d, their standard deviations s_p and s_d, and the correlation coefficient ρ. To randomly draw a job characterized by (p, d), we use the well known fact that the conditional distribution of d given p is also normal with mean $\mu_d+(\rho\, s_d/s_p)(p-\mu_p)$ and standard deviation $s_d(1-\rho^2)^{1/2}$.

To study the effect of the parameters on the solution time, a series of 12 job problems were run using different values for the parameters. Since multiplying the processing times and due dates by an arbitrary constant does not affect the optimal schedule, we can arbitrarily fix one of the parameters—μ_p was set equal to 10 for all problems.

The value of μ_d may be decided by t, the "tardiness" of the shop—a measure of the fraction of the jobs that are completed beyond their due dates. Thus if n is the number of jobs to be scheduled, roughly $n(1-t)$ jobs will be completed by their due dates—the $n(1-t)$th job's completion time $n(1-t)\mu_p$ (on an average) and its due date μ_d (on an average) will be roughly equal. Hence

$$\mu_d=(1-t)(n\mu_p). \tag{6}$$

The standard deviations s_p and s_d can be more conveniently conceptualized by their coefficients of variation s_p/μ_p and s_d/μ_d.

The experimental set-up was a 3^4 factorial design [7], i.e., each of the four parameters could take any one of three values given below (as mentioned already, $n=12$ and $\mu_p=10$).

(a) Mean due date $\mu_d=30,60,90$ $\quad(t=0.75,\ 0.5,\ 0.25)$

(b) Coefficient of variation $s_p/\mu_p=0.20,\ 0.50,\ 0.80$

*This computer program was written by Mr. Nathaniel F. Tarbox, and the author.

(c) Coefficient of variation $s_d/\mu_d = 0.20, 0.50, 0.80$

(d) Coefficient of correlation $\rho = -0.7, 0, 0.7$

For each of the 81 combinations of parameters, 40 problems were solved and the median solution time recorded. The results of the experiments were startling in the sense that the median computation time varied anywhere from 2 msec ($\mu_d = 90$, $s_p/\mu_p = 0.5$, $s_d/\mu_d = 0.5$, $\rho = 0.7$) to 7.021 sec ($\mu_d = 60$, $s_p/\mu_p = 0.5$, $s_d/\mu_d = 0.2$, $\rho = -0.7$) depending on the parameters of the problem. Because of the wide variability of the data, a logarithmic transformation of the median solution times was made. An analysis of variance [7] revealed that the F values for the factors μ_d, s_d/μ_d, and ρ were significant at the 1-percent level. (The first order interaction between ρ and s_p/μ_p was significant at the 5-percent level). A functional relationship between the solution time and the parameters μ_d, s_d/μ_d, ρ, and t was hypothesized by observing the mean solution time for the three values of each parameter. The coefficients were estimated by multiple regression as

$$\ln T = -6.328 + 1.665(1-\rho) + 2.862(1 - s_d/\mu_d) + 26.94\, t - 21.7\, t^2 + \epsilon, \tag{7}$$

where T is the median solution time and ϵ is an error term. The coefficient of multiple correlation was 0.90, and all the regression coefficients were significant at the 1-percent level.

We note that a negative correlation ρ (referred to as the 'perverse' situation by Emmons [5]), and reduced variability of the due dates increases the solution time. To understand the effect of the tardiness factor t, we rewrite (7) as

$$\ln T = 2.022 + 1.665(1-\rho) + 2.862(1 - s_d/\mu_d) - 21.7(t - .62)^2 + \epsilon, \tag{8}$$

where ϵ is an error term. From (8) it is clear that the solution time is maximal when the shop is about 60-percent tardy. It is well known [5] that for a fully 'tardy' shop ($t = 1$), the shortest processing time discipline [2] minimizes total tardiness. For a shop that is least tardy ($t = 0$), it is known that the Earliest Due Date discipline [2] minimizes total tardiness. Since these sequences can be trivially determined, we would expect the solution time to drop drastically both for $t = 0$ and $t = 1$ and this is consistent with (8).

To investigate the effect of the problem size (n), median solution times corresponding $n = 8, 9 \ldots, 50$ were computed based on 40 randomly generated problems for each value of n. The parameters of the problem were set to correspond to the median (of the median) solution time of the 12-job cross-section study just described. These parameters were $\mu_p = 10$, $\mu_d = 5n$ ($t = 0.5$), $s_p/\mu_p = 0.5$, $s_d/\mu_d = 0.80$, and $\rho = 0$ (corresponding to the median solution time of 77 msec for 12-job problems). The median solution times (based on 40 problems) were 11 msec for 8-job problems, and 362 msec for 50-job problems. Thus the hybrid approach clearly avoids the exponential increase of the solution time that is normally the case with dynamic programming problems. A logarithmic regression of the median solution time, T, against the problem size, n, for this combination of parameters was determined to be

$$\ln T = 2.1764 + 0.9154 \ln n, \quad \text{or} \quad T = 8.8 n^{.9154}. \tag{7}$$

Thus the solution time goes up less than linearly with the size of the problem. (The correlation coefficient was 0.8852 and the two regression coefficients were significant at the 0.1-percent level.)

Core limitations prevented the testing of larger problems which would have needed auxiliary storage like drums, discs, or tapes. It is important to note that all the results reported above were median solution times. Even in the 12-job cross-sectional study, 40 out of the 3,240 problems (81×40) were not fully solved because they took longer than 1 min. One may well state that the solution times belong to some infinite variance distribution (For the 12-job problems, the times varied anywhere between 1 msec to more than 1 min).

The solution times for complete enumeration, full dy' mic programming and the Hybrid Algorithm of this paper are compared below (Table 7). The para.. ters used were the 'median' parameters discussed earlier.

TABLE 7. Comparison of Solution Times

Problem size n	Median Solution Times (msec)		
	Complete enumeration	Full dynamic program	Hybrid algorithm
8	28,652	709	11
9	278,438	2,481	67
10	.....	11,931	73
11	.....	41,461	75
12	.....	143,580	77

VI. ACKNOWLEDGMENT

The author wishes to thank Professors Robert S. Kaplan, Paul R. Kleindorfer, David P. Rutenberg, Allan D. Shocker and Gerald L. Thompson for comments on an earlier version of this paper and Mr. Nathaniel F. Tarbox for computational assistance.

REFERENCES

[1] Bellman, R. E., and S. E. Dreyfus, *Applied Dynamic Programming* (Princeton University Press, Princeton, New Jersey, 1962).

[2] Conway, R. W., W. L. Maxwell, and L. W. Miller, *Theory of Scheduling* (Addison-Wesley, Reading, Massachusetts 1967).

[3] Eastman, W. L., "Comments on a paper by Schild and Fredman," Management Science, **11**, 754–5 (1965).

[4] Elmaghraby, S. E., "The One Machine Sequencing Problem with Delay Costs," J. Ind. Eng., **19**, 105–108 (1968).

[5] Emmons, H., "One Machine Sequencing to Minimize Certain Functions of Job Tardiness," Operations Research, **17**, 701–715 (1969).

[6] Held, M., and R. M. Karp, "A Dynamic Programming Approach to Sequencing Problems," SIAM J., **10**, 196–210 (1962).

[7] Kempthorne, O., *The Design and Analysis of Experiments* (John Wiley and Sons, Inc., New York, 1952).

[8] Lawler, E. L., "On Scheduling Problems with Deferral Costs," Management Science, **11**, 280–288 (1964).

[9] Schild, A., and I. J. Fredman, "On Scheduling Tasks with Associated Linear Loss Functions," Management Science, **7**, 280–285 (1961).

[10] Schild, A., and I. J. Fredman, "Scheduling Tasks with Deadlines and Nonlinear Loss Functions," Management Science, **9**, 73–81 (1962).

ON A SEQUENTIAL RULE FOR ESTIMATING THE LOCATION PARAMETER OF AN EXPONENTIAL DISTRIBUTION*

A. P. Basu

Northwestern University†

ABSTRACT

Let us assume that observations are obtained at random and sequentially from a population with density function

$$f(x)=\frac{1}{\sigma}e^{-\left(\frac{x-\mu}{\sigma}\right)}, \qquad x>\mu,\ \mu\geq 0,\ \sigma>0.$$

In this paper we consider a sequential rule for estimating μ when σ is unknown corresponding to the following class of cost functions

$$C_N=A\,|\delta(X_1, \ldots, X_N)-\mu|^p+N,$$

where $\delta(X_1, \ldots, X_N)$ is a suitable estimator of μ based on the random sample $(X_1, \ldots, X_N)$, N is a stopping variable, and A and p are given constants. To study the performance of the rule it is compared with corresponding "optimum fixed sample procedures" with known σ by comparing expected sample sizes and expected costs. It is shown that the rule is "asymptotically efficient" when absolute loss ($p=1$) is used whereas the one based on squared error ($p=2$) is not. A table is provided to show that in small samples similar conclusions are also true.

1. INTRODUCTION

In many physical problems both the one-parameter and the two-parameter exponential distributions serve as useful statistical models. The two-parameter exponential distribution could be a useful model when, for example, a manufacturer may want to determine the "minimum guarantee period" for items he is manufacturing. Thus in problems of life testing a manufacturer of light bulbs or electronic tubes can develop a satisfactory marketing strategy using a two-parameter exponential distribution as model. Another useful case would be trying to predict the life of a cancer patient after a successful operation. In either case, a sequential method of estimation may be desirable if observations are obtained sequentially one at a time. In this paper we consider a sequential rule for estimating the location parameter of the two-parameter exponential distribution when the corresponding scale parameter is unknown.

In section 2 the problem is stated and a sequential rule is given corresponding to a class of cost functions. The operating characteristics of the rules are studied in detail in section 3. Exact distribution of the sample size to terminate the estimating procedure is found for absolute value and squared error loss functions. A table is given showing the expected sample sizes and expected loss functions along with the sample sizes for corresponding optimum fixed sample procedures (where scale pa-

*This research was carried out in part at IBM's Thomas J. Watson Research Center.

†The author is presently with the Department of Mathematics, University of Pittsburgh.

rameter was assumed known). In section 4 a more general rule is stated and its asymptotic properties are proven.

2. STATEMENT OF THE PROBLEM

Let us assume that we obtain observations at random and sequentially from a population with density function . . .

$$f(x)=\frac{1}{\sigma}e^{-\left(\frac{x-\mu}{\sigma}\right)},\ x>\mu,\ \sigma>0. \tag{1}$$

We want to find a sequential rule for estimating the location parameter μ when the scale parameter σ is also unknown. To study the properties of this rule we consider the following cost function:

$$C_N=A|\delta(x_1,\ \ldots,X_N)-\mu|^p+N, \tag{2a}$$

where N denotes the stopping variable indicating the numbers of variables to be observed, $\delta(X_1, \ldots, X_N)$ is an appropriate estimator of μ based on the sample $(X_1, \ldots, X_N)$, and A and p are known positive constants. The first part in (2) is the loss incurred in estimating μ and for $p=1$ and 2 reduces to absolute error and squared error loss function, respectively. The second term in (2) is of course the sampling cost.

Loss functions of this nature are used quite frequently in sample survey, quality control problems, etc., and have a very simple physical interpretation. The difference $\delta(X_1, \ldots, X_N)-\mu$ corresponds to the extent we are in error from true value and is inflated or deflated by the factor A, depending on the relative importance the management wants to attach to it compared to the sampling cost, the second term of (2a) when expressed in units of the amount needed to test a single sampling unit.

Since no fixed sample estimation procedure is available for the above problem, to compare the performance of the proposed sequential rule we shall first consider a particular fixed sample procedure for the following simpler problem. Assume σ is known and consider a fixed random sample of size n from (1). The maximum likelihood estimator of μ, which is also asymptotically unbiased, is then given by the smallest ordered observation $x_{1,n}$. To make a fair comparison with the corresponding sequential estimation problem we choose the sample size for the fixed sample case such that the corresponding expected cost is minimum. We shall denote this sample size as the optimum value n_0 of n. In the sequel we want to compare n_0 with expected sample size of the sequential procedure and the two expected costs. From (2a) the cost function for the fixed sample case will be given by the special form

$$C=A|X_{1,n}-\mu|^p+n \tag{2}$$

$$=A\ (X_{1,n}-\mu)^p+n.$$

It is well known (see, for example, Basu [1]) that

$$u=n(X_{1,n}-\mu)/\sigma \tag{3}$$

follows the exponential distribution with density function

$$g(u)=e^{-u}.$$

It follows that, for fixed n,

$$E[n(X_{1,n}-\mu)/\sigma]^p=\Gamma(p+1). \tag{4}$$

Hence, for fixed n, the expected cost is given by

$$E(C/n)=A(\sigma/n)^p\Gamma(p+1)+n. \tag{5}$$

Considering $E(C/n)$ to be continuous in n, we find the optimum value of n is given by

$$n_0^{p+1}=A\sigma^p p^2\Gamma p \tag{6}$$

and, from (5) and (6), the corresponding minimum cost is

$$E(C/n_0)=n_0(p+1)/p. \tag{7}$$

In case σ is *not known* we may consider the following analogous sequential rule:

Rule R: Sample sequentially obtaining values $(X_1, X_2 \ldots\ldots, X_n, \ldots\ldots)$ and for each $n(n=2, 3, \ldots\ldots)$ compute

$$\hat{\sigma}_n=\left(\sum_{i=1}^{n} X_{i,n}-nX_1,\ n\right)\Big/(n-1) \tag{8}$$

$$=\sum_{i=2}^{n}(n-i+1)(X_{i,n}-X_{i-1,n})/(n-1),$$

from the ordered sample

$$X_{1,n}<X_{2,n}<\ldots\ldots\ldots\ldots\ldots<X_{n,n}.$$

Stop sampling when $N=n\geqslant 2$ for which

$$n^{p+1}\geqq Ap^2\Gamma(p)\cdot\hat{\sigma}_n^p \tag{9}$$

for the first time. That is for which

$$\hat{\sigma}_n\leqq n^{\frac{p+1}{p}}\,[Ap^2\Gamma(p)]^{-1/p}=\sigma(n/n_0)^{\frac{p+1}{p}}=b_n, \text{ say.} \tag{10}$$

Here N denotes the stopping variable.

Since $\hat{\sigma}_n\to\sigma$ with probability one it is clear that the rule R will terminate with probability one. In the next section we shall compare this rule with the corresponding optimum fixed sample procedure with σ known as described before.

3. OPERATING CHARACTERISTICS OF RULE *R*

In this section we study the properties of rule R by finding out the expected sample size needed to terminate the test and the corresponding expected cost function.

Let

$$\frac{\sigma}{2}Z_{i,n}=Y_{i,n}=(n-i+1)(X_{i,n}-X_{i-1,n}),\ (i=2, 3, \ldots, n). \tag{11}$$

Then it is well known that $X_{1,n}$ and $Y_{i,n}$'s (that is, $Z_{i,n}$'s) are mutually independently distributed and $Z_{i,n}$'s are independently and identically distributed each following the χ^2-distribution with two degrees of freedom.

Denoting $P(N=n)$ by p_n, we have under R the cost function

$$C_N=(X_{1,N}-\mu)+N \tag{12}$$

with expectation

$$\begin{aligned} E(C_N|R) &= \sum_{n=2}^{\infty} P_n E(C/n) \\ &= E(N)+A\sigma^p\Gamma(p+1)E(N^{-p}) \\ &= E(N)+\frac{n_0^{p+1}}{p}E(N^{-p}). \end{aligned} \tag{13}$$

From (10)

$$p_n=P(\hat{\sigma}_2>b_2,\ \hat{\sigma}_3>b_3,\ \ldots,\ \hat{\sigma}_{n-1}>b_{n-1},\ \hat{\sigma}_{n\leqq b_n}) \tag{14}$$

$$=P(S_1>a_2,\ S_2>a_3,\ \ldots,\ S_{n-2}>a_{n-1},\ S_{n-1}\leqq a_n) \tag{15}$$

where

$$S_{n-1}=\sum_{i=2}^{n} Z_{i,n} \qquad \text{and} \qquad a_n=\frac{2(n-1)}{\sigma}b_n \qquad (n\geqslant 2)$$

$$=2(n-1)\left(\frac{n}{n_0}\right)^{\frac{p+1}{p}}$$

J. E. Moyal has given an algorithm for computing p_n and it is described in Robbins [3], which may be stated briefly as follows.

Let $m=n-1$, $a_m=2(m-1)(m/n_0)^{\frac{p+1}{p}}$ $\qquad (m=1, 2\ \ldots\ .)$

$h_1(\cdot)=1$ and $C_1=1$.

Compute recursively

$$h_m(a_n)=\sum_{j=1}^{m-1}\frac{(a_n-a_m)^j}{j!}h_{m-j}(a_m), \qquad \begin{matrix} m=2, 3\ \ldots\ . \\ n=m+1, m+2\ \ldots\ . \end{matrix}$$

$$C_m=e^{-a_m}\sum_{j=1}^{m-1}h_{m-j}(a_m),\ C_1=1, \quad (m=2, 3, \ldots\ .)$$

Then,

$$p_m=C_m-C_{m+1} \qquad (m=1, 2, \ldots\ .).$$

Using the above algorithm we have computed p_n for $n_0=2(1)50$ and corresponding cost function when $p=1$ and 2. For the sake of comparison A and σ are so chosen that each n_0(with $p=1, 2$) is the optimum sample size in the corresponding "fixed sample" case. Some of these values are given in Table I along with corresponding optimum "fixed sample size" n_0 and cost.

TABLE I. *Optimum or Expected Sample Sizes and Expected Costs for Single Sample and Sequential Estimation Rules*

n_0	E(N/R)	E(C/n_0)	E(C/R)	n_0	E(N/R)	E(C/n_0)	E(C/R)	n_0	E(N/R)	E(C/n_0)	E(C/R)
$p=.25$				$p=.50$				$p=1.00$			
2	2.135	10.000	10.031	2	2.135	6.000	6.036	2	2.136	4.000	4.046
3	2.855	15.000	15.079	3	2.634	9.000	9.172	3	2.488	6.000	6.314
4	3.730	20.000	20.111	4	3.285	12.000	12.377	4	2.947	8.000	8.903
5	4.638	25.000	25.113	5	4.021	15.000	15.543	5	3.481	10.000	11.608
6	5.542	30.000	30.107	6	4.806	18.000	18.651	6	4.069	12.000	14.319
7	6.435	35.000	35.102	7	5.615	21.000	21.712	7	4.699	14.000	16.977
8	7.320	40.000	40.100	8	6.435	24.000	24.743	8	5.360	16.000	19.559
9	8.199	45.000	45.103	9	7.257	27.000	27.758	9	6.044	18.000	22.057
10	9.075	50.000	50.107	10	8.079	30.000	30.765	10	6.744	20.000	24.478
11	9.950	55.000	55.114	11	8.897	33.000	33.771	11	7.456	22.000	26.830
12	10.824	60.000	60.122	12	9.712	36.000	36.778	12	8.176	24.000	29.124
13	11.697	65.000	65.130	13	10.523	39.000	39.788	13	8.902	26.000	31.371
14	12.570	70.000	70.140	14	11.332	42.000	42.801	14	9.630	28.000	33.582
15	13.442	75.000	75.149	15	12.137	45.000	45.817	15	10.359	30.000	35.765
16	14.314	80.000	80.159	16	12.941	48.000	48.837	16	11.089	32.000	37.927
17	15.186	85.000	85.169	17	13.743	51.000	51.858	17	11.819	34.000	40.073
18	16.057	90.000	90.180	18	14.544	54.000	54.882	18	12.548	36.000	42.209
19	16.928	95.000	95.190	19	15.344	57.000	57.908	19	13.275	38.000	44.338
20	17.800	100.000	100.201	20	16.143	60.000	60.935	20	14.002	40.000	46.462
21	18.671	105.000	105.212	21	16.941	63.000	63.963	21	14.727	42.000	48.583
22	19.542	110.000	110.223	22	17.738	66.000	66.993	22	15.451	44.000	50.703
23	20.413	115.000	115.234	23	18.535	69.000	70.023	23	16.173	46.000	52.821
24	21.284	120.000	120.245	24	19.332	72.000	73.054	24	16.895	48.000	54.940
25	22.155	125.000	125.256	25	20.128	75.000	76.086	25	17.165	50.000	57.059
$p=1.50$				$p=2.00$				$p=2.50$			
2	2.138	3.333	3.388	2	2.139	3.000	3.063	2	2.141	2.800	2.871
3	2.437	5.000	5.454	3	2.412	4.500	5.108	3	2.397	4.200	4.981
4	2.821	6.667	8.152	4	2.757	6.000	8.197	4	2.719	5.600	8.696
5	3.266	8.333	11.263	5	3.154	7.500	12.194	5	3.086	7.000	14.138
6	3.758	10.000	14.616	6	3.593	9.000	16.937	6	3.492	8.400	21.312
7	4.288	11.667	18.086	7	4.066	10.500	22.267	7	3.929	9.800	30.153
8	4.850	13.333	21.588	8	4.569	12.000	28.043	8	4.394	11.200	40.563
9	5.437	15.000	25.062	9	5.096	13.500	34.150	9	4.882	12.600	52.434
10	6.045	16.667	28.473	10	5.645	15.000	40.493	10	5.391	14.000	65.664
11	6.671	18.333	31.802	11	6.212	16.500	47.003	11	5.918	15.400	80.157
12	7.309	20.000	35.040	12	6.795	18.000	53.626	12	6.461	16.800	95.837
13	7.959	21.667	38.187	13	7.390	19.500	60.326	13	7.018	18.200	112.639
14	8.617	23.333	41.246	14	7.995	21.000	67.079	14	7.586	19.600	130.516
15	9.282	25.000	44.226	15	8.610	22.500	73.868	15	8.164	21.000	149.433
16	9.951	26.667	47.134	16	9.232	24.000	80.687	16	8.750	22.400	169.367
17	10.624	28.333	49.979	17	9.860	25.500	87.533	17	9.344	23.800	190.305
18	11.299	30.000	52.770	18	10.492	27.000	94.407	18	9.944	25.200	212.243
19	11.976	31.667	55.516	19	11.129	28.500	101.312	19	10.549	26.600	235.183
20	12.654	33.333	58.223	20	11.768	30.000	108.253	20	11.158	28.000	259.130
21	13.332	35.000	60.900	21	12.410	31.500	115.237	21	11.770	29.400	284.095
22	14.011	36.667	63.550	22	13.053	33.000	122.270	22	12.386	30.800	310.090
23	14.689	38.333	66.180	12	13.697	34.500	129.358	23	13.003	32.200	337.131
24	15.367	40.000	68.793	24	14.343	36.000	136.506	24	13.623	33.600	365.232
25	16.045	41.667	71.393	25	14.988	37.500	143.721	25	14.243	35.000	394.409

TABLE I. *Optimum or Expected Sample Sizes and Expected Costs for Single Sample and Sequential Estimation Rules*—Continued

n_0	E(N/R)	E(C/n_0)	E(C/R)
		$p=3.00$	
2	2.141	2.667	2.744
3	2.387	4.000	4.980
4	2.693	5.333	9.584
5	3.041	6.667	17.253
6	3.424	8.000	28.518
7	3.836	9.333	43.775
8	4.275	10.667	63.322
9	4.736	12.000	87.388
10	5.217	13.333	116.162
11	5.716	14.667	149.808
12	6.230	16.000	188.485
13	6.758	17.333	232.345
14	7.299	18.667	281.551
15	7.849	20.000	336.270
16	8.409	21.333	396.679
17	8.977	22.667	462.965
18	9.551	24.000	535.323
19	10.132	25.333	613.957
20	10.717	26.667	699.075
21	11.307	28.000	790.893
22	11.900	29.333	889.630
23	12.496	30.667	995.506
24	13.095	32.000	1108.744
25	13.696	33.333	1229.568

It is seen that for both $p=1$ and $p=2$ $E(N|R)$ is smaller than the corresponding optimum sample size n_0; however, the ratio of expected costs $\frac{E(C|R)}{E(C|n_0)}$ seems to increase at a much faster rate when $p=2$ than when $p=1$.

Thus it is clear from Table I that while the rule is a very satisfactory one when the absolute loss is used, the case for squared error loss is quite unsatisfactory. In fact, it seems the cost function in this case is becoming infinitely large as $n_0 \to \infty$ (that is as $\sigma \to \infty$). In the next section we assert that this will be the case in general by considering a slightly more general rule *GR* and considering the asymptotic properties of this rule.

4. ASYMPTOTIC PROPERTIES OF *GR*

In this section we consider a more general rule which may be stated as follows:

Rule GR: As in Rule *R* compute $\hat{\sigma}_n$ and stop sampling for the first $N=n \geq m$ for which (9) holds, where m is a fixed integer greater than or equal to 2.

It is clear that *R* is a special case of *GR* when $m=2$. For *GR* all the previous equations and expressions remain unchanged except for obvious modifications such as

$$p(N=n)=0 \qquad \text{for } N<m.$$

To evaluate the asymptotic performance of *GR*, following Starr [4], let us define for each σ the "effi-

ciency" of *GR* with respect to the fixed sample procedure as

$$\eta(\sigma) = \frac{\text{expected cost under } GR}{\text{expected cost under fixed sample procedure}}$$

$$= \frac{E(C_N|R)}{E(C|n_0)} = \frac{\sum_{n=m}^{\infty} p_n E(c/n)}{n_0(p+1)/p}$$

$$= \frac{E(N) + n_0^{p+1} E(N^{-p})/p}{n_0(p+1)/p}$$

$$= \frac{1}{p+1} [pE(N/n_0) + n_0^p E(N^{-p})]. \tag{16}$$

We shall compare the performance of the proposed sequential rule with the optimum fixed sample procedure by showing that for some values of the rule *GR* is *asymptotically efficient.*

The following theorems will establish the asymptotic properties of *GR* as $\sigma \to \infty$ and show under what condition *GR* is "asymptotically efficient," that is

$$\lim_{\sigma \to \infty} \eta(\sigma) = 1. \tag{17}$$

THEOREM 1: With n_0 defined by (6) and with N defined in *GR*

$$\cdot p \lim_{\sigma \to \infty} \frac{N}{n_0} = 1. \tag{18}$$

PROOF: Follows as a corollary to Lemma 1 of Chow and Robbins [2].

THEOREM 2: For $w > o$ fixed

$$\lim_{\sigma \to \infty} E\left(\frac{N}{n_0}\right)^w = 1. \tag{19}$$

PROOF: Same as Theorem 2 of Ref. [4].

THEOREM 3: For $w > o$ fixed

$$\lim_{\sigma \to \infty} n_0^w E N^{-w} = 1 \qquad \text{if } m > 1 + pw/(p+1),$$

$$= \{2(m-1)m^{(p+1)/p}\}^{m-1}, \qquad \text{if } m = 1 + \frac{pw}{p+1}$$

$$= \infty, \qquad \text{if } m < 1 + \frac{pw}{p+1}.$$

PROOF: Let $\alpha = (1-\epsilon)^{1/w} n_0$, $\beta = (1+\epsilon)^{1/w} n_0$. Now,

$$EN^{-w} \geq m^{-w} P(N = m) + \beta^{-w} P(m < N \leq \beta)$$

$$= m^{-w} P(S_{m-1} \leq a_m) + \beta^{-w} P(m < N \leq \beta),$$

$$= m^{-w} \frac{1}{\Gamma(m-1)} \int_0^{a_m} e^{-x} x^{m-2} dx + \beta^{-w} p(m < N \leq \beta)$$

$$\geq \frac{m^{-w}}{\Gamma m} e^{-a_m} a_m^{m-1} + \beta^{-w} P(m < N \leq \beta).$$

Since

$$\lim_{\sigma\to\infty} a_n = 0 \text{ for any fixed } n,$$

and

$$\lim_{\sigma\to\infty} P(m < N \leqslant \beta) = 1,$$

we have

$$\liminf_{\sigma\to\infty} n_0^w E N^{-w} \geqslant \frac{m^{-w}}{\Gamma m} \lim_{\sigma\to\infty} (n_0^w a_m^{m-1}) + 1 - \delta, \text{ say}$$

where

$$0 < \delta = \delta(\epsilon) < 1. \tag{21}$$

Also,

$$EN^{-w} \leqslant m^{-w} P(N=m) + m^{-w} P(m < N \leqslant \alpha) + \alpha^{-w} P(N \geqslant \alpha) = t_1 + t_2 + t_3, \text{ say.} \tag{22}$$

But,

$$\begin{aligned} t_1 = m^{-w} P(S_{m-1} \leqslant a_m) &= \frac{m^{-w}}{\Gamma m - 1} \int_0^{a_m} e^{-x} x^{m-2} dx \\ &\leqslant \frac{m^{-w}}{\Gamma m - 1} \int_0^{a_m} x^{m-2} dx \\ &= \frac{m^{-w}}{\Gamma m} a_m^{m-1}. \end{aligned} \tag{23}$$

Since

$$P(N=n) \leqslant P(S_{n-1} \leqslant a_n) \qquad (n \geqslant m),$$

we get

$$t_2 \leqslant m^{-w} \sum_{m<n\leqslant\alpha} \frac{1}{\Gamma n - 1} \int_0^{a_n} e^{-x} x^{n-2} dx \leqslant m^{-w} \sum_{m<n\leqslant\alpha} \frac{a_n^{n-1}}{\Gamma n}.$$

But it can be easily checked that $\left\{\frac{a_n^{n-1}}{n}\right\}$ is an increasing sequence.

Hence,

$$t_2 \leqslant m^{-w} \frac{a_r^{r-1}}{\Gamma r} (r-m), \tag{25}$$

where $r = [\alpha]$, the greatest integer contained in n. Also, note that, for large r

$$\frac{n_0^w \cdot a_r^{r-1}(r-m)}{\Gamma r} \simeq n_0^w \cdot a_r^{r-1}/\Gamma(r-1) = 0(n_0^{w-(r-1)(p+1)/p}/\Gamma(r-1)), \tag{26}$$

which $\to 0$ as $\sigma \to \infty (n_0 \to \infty)$. Thus, from (22), (23), (24), (25), and (26) we have

$$\limsup_{\sigma\to\infty} n_0^w E N^{-w} \leqslant \frac{m^{-w}}{m} \lim_{\sigma\to\infty} (n_0^w a_m^{m-1}) + 0 + 1 + \tau, \tag{27}$$

where

$$(0 < \tau = \tau(\epsilon) < 1).$$

Combining (21) and (27), we get

$$\lim_{\sigma\to\infty} n_0^w E N^{-w} = \frac{m^{-w}}{\Gamma m} \lim_{\sigma\to\infty} (n_0^w a_m^{m-1}) + 1. \tag{28}$$

Now, from (16),

$$n_0^w a_m^{m-1} = \{2(m-1)\}^{m-1} \cdot m^{(p+1)(m-1)/p} \cdot n_0^{w-(p+1)(m-1)/p}. \tag{29}$$

As $\sigma \to \infty \iff n_0 \to \infty$, the theorem follows from (29).

Theorems 1, 2, and 3 lead to the following interesting corollary.

COROLLARY: For $p > 0$ fixed

$$\lim_{\sigma\to\infty} \eta(\sigma) = 1 \qquad \text{for } m > 1 + pw/(p+1), \tag{30}$$

$$= \frac{p}{p+1}\left\{1 + (2(m-1)m^{(p+1)/p})\right\} \qquad \text{for } m = 1 + \frac{pw}{p+1}$$

$$= \infty \qquad \text{for } m < 1 + \frac{pw}{p+1}.$$

It follows that in the case of the sequential distribution, the best sequential rule is obtained when $p \leq 1$ and $m \geq 2$. The above theorems also explain Table I. In fact, for large σ it seems $\eta(\sigma)$ is a monotone decreasing function of p. Thus any reasonable loss function with $p \leq 1$ may be chosen in the case of exponential distribution.

REFERENCES

[1] Basu, A. P., "On Some Tests of Hypotheses Relating to the Exponential Distribution When Some Outliers Are Present," *J. Am. Statist. Assoc.*, **60**, 548–559 (1965).

[2] Chow, Y. S. and H. Robbins (1965). "On the Asymptotic Theory of Fixed Width Sequential Confidence Intervals for the Mean," Ann. Math. Statist., **36**, 457–462 (1965).

[3] Robbins, H. (1958). "Sequential Estimation of the Mean of a Normal Population," *Probability and Statistics* (Harald Cramer Volume) (Uppsala: Almquist and Wiksell, 1958).

[4] Starr, Norman, "On the Asymptotic Efficiency of a Sequential Procedure for Estimating the Mean," Ann. Math. Statist., **37**, 1173–1185 (1966).

A GRAPH THEORETIC INTERPRETATION OF THE SUFFICIENCY CONDITIONS FOR THE CONTIGUOUS-BINARY-SWITCHING (CBS)-RULE*

Salah E. Elmaghraby

North Carolina State University
Raleigh, N.C.

ABSTRACT

A sufficient condition for the optimality of the *CBS*-rule due to W. Smith is given a graphic interpretation in terms of 'convex' graphs. A convex graph is uniquely constructed (except for a homomorphism), and has the property that the optimum is achieved from any starting point.

I. INTRODUCTION

The problem of concern to us here is that of scheduling n independent jobs on $M \geq 1$ facilities in series, in which all jobs possess the same route (i.e., we are dealing with a 'flow shop'). The processing time of each job on each facility is given and may include any setup time. The objective is to minimize a given function, f, of the job completion times.

Several such scheduling problems have been successfully resolved by application of the 'contiguous-binary-switching' rule (the *CBS*-rule for short), to be described presently. Examples are: the scheduling of n jobs on two processors in series to minimize the total 'makespan,' treated by Johnson [2]; the scheduling of n jobs on a single facility to minimize a weighted sum of completion times, treated by Smith [4]; the scheduling of n jobs on two machines in series with arbitrary start- and stop-lags, treated by Mitten [3]; among others.

Simply stated, the *CBS*-rule essentially says the following: start with any arbitrary sequence, say Q_1, where a sequence is defined to be a permutation of the numbers $1,2, \ldots, n$ on each machine; and let Q_1 be the matrix whose mth row is given by $(q_1^{(m)}, q_2^{(m)}, \ldots, q_n^{(m)})$, where $q_i^{(m)}$ denotes the job in sequence position i on machine m; $m = 1,2, \ldots, M$. Test the reversal of the order of any contiguous pair of jobs, say $q_i^{(m)}$ and $q_{i+1}^{(m)}$, for its impact on a function g defined on the ordered pair $(q_i^{(\cdot)}, q_{i+1}^{(\cdot)})$. If the switch decreases g, perform the switch (recall we are minimizing f) to obtain a new sequence Q_2; otherwise do not perform the switch. Repeat the test on the new sequence, moving to sequences $Q_3, Q_4, \ldots, Q^*$. The sequence Q^* in which the test indicates that no switching of *contiguous* jobs yields an improvement in g is the *optimal* sequence. Note that in any iteration of the procedure one performs at most $(n-1)M$ tests of the type indicated.

It is well-known that the *CBS*-rule is not a universally applicable procedure, in the sense that one can easily construct examples in which a sequence $\hat{Q}$, say, is obtained in which no switching of any pair of *contiguous* jobs yields an improvement in the objective function $f(\hat{Q})$, and yet switching two (or more) *noncontiguous* jobs yields an improvement. Consequently, $\hat{Q}$ is a *local* optimal sequence, but not a global one.

*This research was partially supported by NSF Grant No. GK-2647.

Equally well-known is the fact that the *CBS*-condition is a *necessary*, but not sufficient condition for optimality (for otherwise, by performing the indicated binary switch one would achieve a better sequence).

The need for sufficient conditions for optimality was realized quite early in the development of scheduling theory. In particular, Smith [4] provided and proved the following sufficient condition (slightly reworded to conform to our notation):

Smith's Sufficiency Theorem: A sufficient condition that $f(Q^*) \leq f(Q)$ for all permutations Q of the n jobs is that:

(a) There exists a real-valued function g of ordered pairs of elements *s.t.* if Q is any permutation and Q' the permutation obtained from Q by interchange of the ith and the $(i+1)$st elements, then

$$g(q_i, q_{i+1}) \leq g(q_{i+1}, q_i) \Rightarrow f(Q) \leq f(Q');$$

(b) Q^* is *s.t.* job i precedes job j if

$$g(i, j) \leq g(j, i).$$

In the following we give a graph-theoretic interpretation of this Theorem in the hope that such treatment will shed more light on the significance of the conditions (a) and (b), and possibly lead to either modified conditions or other procedures in the cases in which the *CBS*-rule is inapplicable.

The Graphic Representation

We shall construct two graphs, called H and G, in the following fashion.

Represent each possible sequence by a node (or vertex). Denote the set of nodes by N. (Thus, in the case of a single facility there are $|N| = n!$ nodes; and in the case of two facilities there are $|N| = (n!)^2$ nodes; etc.) Connect node Q_1 to node Q_2 by an *undirected* arc (Q_1, Q_2) if Q_2 can be obtained from Q_1 by a *CBS*. (Obviously, in this case Q_1 can be obtained from Q_2 by the reverse *CBS*.) Nodes Q_1 and Q_2 are said to be *adjacent* to each other. More formally, if N is a set of points, we define the mapping A on the unordered pair of points Q_1 and Q_2 as follows:

$$A = \{(Q_1, Q_2) \in N \times N : Q_1 \text{ and } Q_2 \text{ are mutually adjacent}\}. \tag{1}$$

The resulting graph $H = (N, A)$ is called the *connecting* graph; which is:

(1) Connected and undirected: since it is easy to show that any sequence Q can be obtained from any other sequence Q' by a series of *CBS*'s.

(2) Each node is of rank $(n-1)M$; i.e., there are $(n-1)M$ arcs incident on each node, since on any machine there are exactly $n-1$ possible *CBS*'s and there are M such machines.

(3) The shortest path between any pair of nodes is of length 1 (if the two nodes are adjacent), and the longest path is of the order $(n!)^M$, though slightly less than the total number of nodes in N.

Figure 1 illustrates the connecting graph for $n = 4$ jobs on a single facility. There are, in all, $|N| = 4! = 24$ different sequences which have been enumerated in the table of Fig. 1.

The construction of the graph G proceeds as follows.

The objective function, f (to be minimized) maps each node into the real line. Consequently, by

Feasible Sequences:

1	1,2,3,4	7	2,1,3,4	13	3,1,2,4	19	4,1,2,3
2	1,2,4,3	8	2,1,4,3	14	3,1,4,2	20	4,1,3,2
3	1,3,2,4	9	2,3,1,4	15	3,2,1,4	21	4,2,1,3
4	1,3,4,2	10	2,3,4,1	16	3,2,4,1	22	4,2,3,1
5	1,4,2,3	11	2,4,1,3	17	3,4,1,2	23	4,3,1,2
6	1,4,3,2	12	2,4,3,1	18	3,4,2,1	24	4,3,2,1

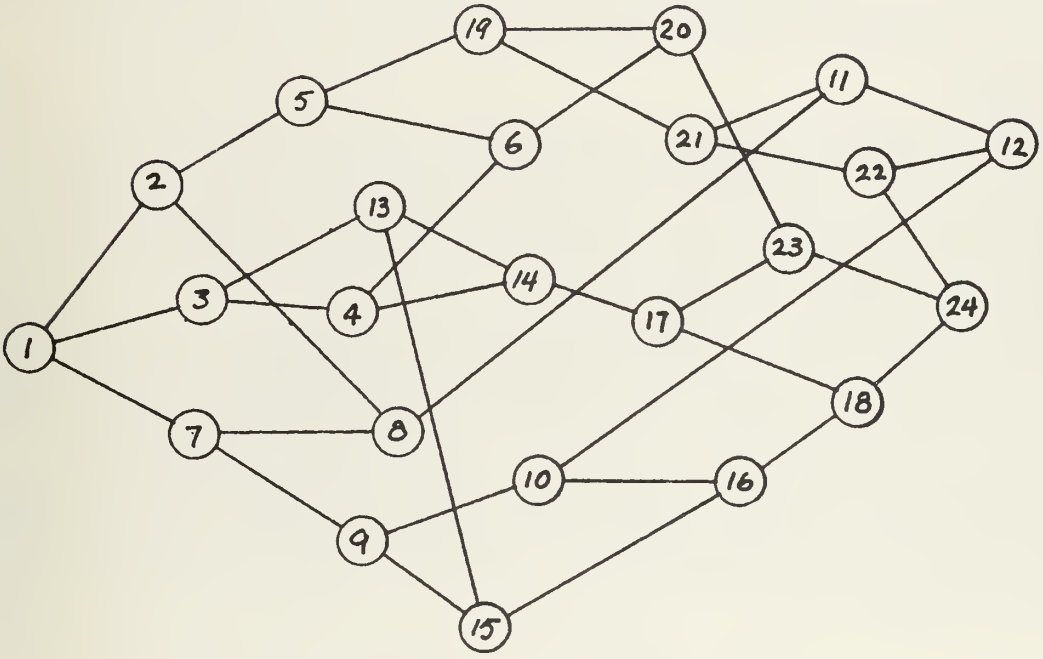


FIGURE 1. The connecting graph $H=(H, A)$

the complete ordering property of the real line, each node Q can be put into one of three relations to each *adjacent* node Q':

either (i) $f(Q)<f(Q')\Rightarrow$ the arc is directed from Q' to Q

or (ii) $f(Q)>f(Q')\Rightarrow$ the arc is directed from Q to Q'

or (iii) $f(Q)=f(Q')\Rightarrow$ leave the arc undirected.

This construction leads to a special kind of "preference graph," in particular, one in which is defined a relation Γ' on the nodes N as follows:

$$\Gamma'=\{(Q,Q')\in N\times N: f(Q')<f(Q) \text{ and } (Q,Q')\in A\}. \tag{2}$$

Thus any pair of vertices Q and Q' are joined by an arc *directed* from Q to Q' iff $f(Q') < f(Q)$ and the arc (Q, Q') exists in the connecting graph H (i.e., Q can be obtained from Q' by a *CBS*). Furthermore, there is also defined a relation Γ'' on the nodes N as follows:

$$\Gamma'' = \{(Q, Q') \in N \times N: f(Q) = f(Q') \text{ and } (Q, Q') \in A\}. \tag{3}$$

Thus any pair of vertices Q and Q' are joined by an *undirected* arc iff $f(Q) = f(Q')$ and the arc $(Q, Q') \in A$. Finally, we let,

$$\Gamma = \Gamma' \cup \Gamma''. \tag{4}$$

The graph $G = (N, \Gamma)$ is called the preference graph and possesses the following properties:

(1) Antisymmetric under Γ' and symmetric under Γ''.

(2) Has no self-loops.

Unlike ordinary preference graphs, the graph G is not transitive (i.e., if $(Q, Q'), (Q', Q'') \in \Gamma$, then (Q, Q'') need not be in Γ except when $(Q, Q'') \in A$, in which case the arc (Q, Q'') would be in Γ).

'Convex' Graphs†

We now wish to introduce the concept of 'convex directed graph', (or CDG):

Definition: A directed graph is said to be 'convex' if its nodes are partitioned into two subsets B and R *s.t.* the following three conditions are satisfied: (i) if B is not empty then every node in B has at least one arrow out of it in the direction of improving f; (ii) the subset R is not empty; and (iii) $f(Q) =$ minimum for all $Q \in R$.

As will become evident shortly, the rationale for the name 'convex' stems from the fact that, in such a graph, one can start in any node in the subset B and proceed monotonically in the direction of improving f until the subset R is reached, at which time no improvement is possible. Hence, any node (i.e., sequence) in R is optimal. (This is reminiscent of a convex function over a bounded set.)

It is easy to see that a desire to minimize $f(Q)$ is equivalent to selecting a node $Q^* \in N$ over which no other node is preferred, i.e., all arcs incident on Q^* are generated by either Γ', in which case they are of the form (Q, Q^*), and directed *into* Q^*; or Γ'', in which case they are undirected. Stated differently, Q^* is optimal if

$$\Gamma(Q^*) \neq \phi \Rightarrow (Q^*, Q) \in \Gamma'', \text{ undirected.}$$

Hence Q^* has no 'successor' under Γ', in the sense of absence of *directed* arcs *out* of Q^*. Any node having no successors under Γ' is called a *receiver* of the graph. The set

$$R = \{Q^* \in N: \Gamma(Q^*) \neq \phi \Rightarrow (Q^*, Q) \in \Gamma''\}$$

is called the *set of receivers* of a graph $G = (N, \Gamma)$. Thus the problem of finding a minimizing sequence reduces to finding any member of R; and the existence of an optimum is equivalent to the requirement that R be non-empty; which is condition (ii) of the definition of *CDG*.

In the case of scheduling n jobs on M machines, the finiteness of the space of feasible sequences and the boundedness of the function f guarantee the non-emptiness of R. Thus, *condition* (ii) *of the*

†The word *convex* was used by F. Glover ("Maximum Matching in a Convex Bipartite Graph," Nav. Res. Log. Quart., **14**, 3, 313–316, Sep 1967) in the context of bipartite graphs. His definition of convexity in graphs bears no relationship to ours.

definition of a CDG *is always satisfied* in the class of scheduling problems of interest to us here; i.e., in the graph $G=(N, \Gamma)$, the subset R is nonempty.

It is equally easy to demonstrate that *invoking condition* (i) *of the definition* of a *CDG*, we immediately conclude that $f(Q_r^*)=$constant for all nodes $Q_r^* \epsilon R$, $r=1, \ldots, |R|$. For let Q_1^* and Q_2^* be two nodes in R; then since

and

$$\text{either} \quad \Gamma(Q_1^*)=\phi \quad \text{or} \quad (Q_1^*, Q)\epsilon\Gamma'', Q\epsilon R,$$
$$\text{either} \quad \Gamma(Q_2^*)=\phi \quad \text{or} \quad (Q_2^*, Q')\epsilon\Gamma''; Q'\epsilon R;$$

then any two vertices in R are either not adjacent or are connected by an *undirected* arc. If Q_1^* and Q_2^* are adjacent then, *a fortiori*, $f(Q_1^*)=f(Q_2^*)$. And if they are not adjacent then a simple contradiction establishes that they must be of equal value (for otherwise either Q_1^* or Q_2^* must have a 'successor').

Thus we conclude that in G condition (i) is sufficient to establish condition (iii). Since we have already established that condition (ii) is always satisfied, we conclude that:

ASSERTION 1: The graph $G=(N, \Gamma)$ is convex iff condition (i) in the definition is satisfied.

Now let Q be an arbitrary node *not* in R, and let Q_r^* be some node in R. We remark that, by construction, $f(Q) \not\geq f(Q_r^*)$. We have:

ASSERTION 2: If G is convex then there exists at least one directed path from Q to some $Q_r^* \epsilon R$.

PROOF: The proof is by appeal to the construction of G. Suppose, to start with, that R is a singleton set; then Q^* must be the only node in R and Q^* must be 'reachable' from any other node $Q \notin R$. For, among all nodes adjacent to Q there must exist at least one node, say Q', *s.t.* $f(Q')$ is $<f(Q)$; by the assumption of convexity of G. If $Q' \equiv Q^*$, we are done; otherwise, the same argument applies to node Q' which must possess an adjacent node Q'', say *s.t.* $f(Q'')<f(Q')$. Proceeding in this manner we must terminate at Q^*, otherwise an obvious contradiction would result. This establishes the assertion in the case of a unique optimum Q^*.

The case of multiple optima, $Q_1^*, Q_2^*, \ldots, Q_{|R|}^*$, follows similar reasoning, the difference being that there may be no directed path between any two arbitrarily chosen nodes $Q \notin R$ and $Q_r^* \epsilon R$.

We turn next to the 'bad' end of the graph G. By the finiteness of G there must exist at least one node in B whose arcs are leading out of it in the direction of improving f. From the point of view of the objective function, f, these nodes are the 'worst' sequences. Call the collection of such nodes the subset $W(\subset B)$. We may make the

ASSERTION 3: If G is convex then W is nonempty, and $f(Q)=$constant for all $Q \epsilon W$. The proof is similar to that of Assertion 2 and therefore will not be repeated. Finally, the above development leads to the

THEOREM: The sufficient conditions of Smith are equivalent to the requirement that the preference graph $G=(N, \Gamma)$ satisfies condition (i) of the Definition. Figure 2 illustrates such a *CDG* constructed for the example of Fig. 1 with

$$f(Q)=\sum_j T_j,$$

where $T_j=\sum_{i \leq j} t_i$, the completion time of job j in the sequence Q. The processing times of the four jobs, together with $f(Q)$ for all possible 24 sequences are also given in Fig. 2.

i	1	2	3	4
t_i	1	3	4	6
c_i	4	3	1	5

Q	1	2	3	4	5	6	7	8	9	10	11	12
f(Q)	94	80	103	106	83	92	103	89	118	137	108	123

Q	13	14	15	16	17	18	19	20	21	22	23	24
f(Q)	118	121	127	146	140	149	102	111	111	126	126	135

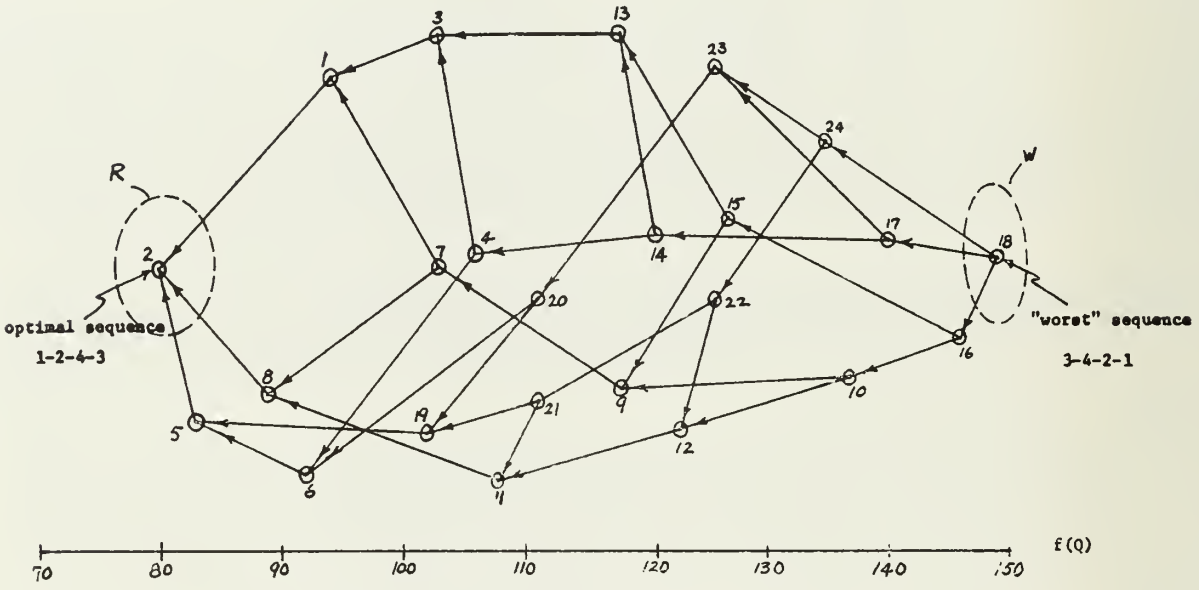


FIGURE 2. Convex Directed Graph $G=(N, \Gamma')$ of Figure 1.

REFERENCES

[1] Berge, C., *The Theory of Graphs and its Applications* (Wiley, New York, 1962).

[2] Johnson, S. M., "Optimal Two- and Three-Stage Production Schedules with Setup Times Included," Nav. Res. Log. Quart. **1**, 61–68 (1954).

[3] Mitten, L. G., "A Scheduling Problem," J. Ind. Eng., Vol. 10, No. 2 (1959).

[4] Smith, W. E., "Various Optimizers for Single-Stage Production," Nav. Res. Log. Quart. **3**, 59–66 (1956).

POLITICAL GAMES

Guillermo Owen

Rice University
Houston, Texas

ABSTRACT

A modification of the Shapley value is suggested, which takes into account the fact that (due to personal affinities among the players) certain coalitions are more easily formed than others. This is done by assigning to each player a point in space, and looking at the distances between pairs of points. The method seems to be especially applicable to voting games among political parties (in, e.g., parliaments), and, for such games, gives a value which is considerably easier to compute than the usual Shapley value. Some examples are considered.

It is a well known fact that, in most games, the players do not behave as one would expect from an abstract study of the game. That is, the characteristic function, or even the normal or extensive forms of the games are not sufficient to determine the coalitions which will form, since these depend to a large extent, on personal affinities of the players.

An obvious example appears in political situations. We may consider a parliament as a weighted voting game, in which the "players" are the parties represented in parliament (we assume here that each party leader can control, to some degree at least, the manner in which his parliamentarians will vote; otherwise it is not really a homogenous party, and should be decomposed into homogenous groupings.) It is easily observed, now, that there are certain parties, whose voting power in the "abstract" game (the Shapley value) is only slightly related to the "payoff" (say, the number of cabinet posts) received. We may, for instance, consider the French Communist party, which through the days of the Fourth Republic was consistently the largest single party, and yet never managed to enter into a government coalition. The reason for this is simply that the other parties were never willing to join it in a coalition: thus, the coalitions including this party were, if not impossible, at least highly improbable: improbable enough that no such coalition ever formed.

We wish to give a modification of the Shapley value: one which will take into account the fact that certain coalitions are considerably more probable than others.

As Shapley points out in [3], his value can be obtained in the following manner: Consider all possible $n!$ orderings of the n players in a game, and assign the probability $1/n!$ to each of these. For a given ordering, $\rhd$, we let $S(i, \rhd)$ be the set of players which precede player i under the ordering $\rhd$. Then the Shapley value $\phi[v]$ of the game is defined by

$$\phi_i[v] = E[v(S(i, \rhd) \cup \{i\}) - v(S(i, \rhd))], \tag{1}$$

where E denotes expected value under the given randomization scheme. It should be pointed out here that expression (1) assures, whatever randomization scheme may be used, that the value ϕ will be additive, that a carrier T of the game v will receive the amount $v(T)$, and, finally, for super-additive games, that the value ϕ will be an imputation.

We note that for the Shapley value, all the orderings of the players are assigned the same probability. The reason for this is that, as Shapley is careful to explain, nothing other than the characteristic function is assumed known about the game. It is assumed that any differences among the players can be explained in terms of the characteristic function, or else do not enter into the calculation of the value; this is the meaning of the symmetry axiom (see Axiom A1 in [3]). (In a stricter explanation Shapley explains that the value is a function *only* of the characteristic function, thought of as a set function.) There is indeed much sense in this for personal, psychological differences among the players can be extremely difficult to measure, and it is often necessary to analyze a game before any study of players can be made. Yet the psychological differences exist, and it should be admitted that at least in the case of political situations, some knowledge exists of the affinities among the several players of the game. We know, from experience, that right and left exist, and that a right-wing party and a left-wing party will generally do anything rather than enter into a coalition with each other, even if "abstract" game theory suggests that this is their best course of action. One concludes that orderings of the parties which place two such parties close together, and in a pivotal position, should be assigned low probability.

Granted that not all orderings of the players should be assigned the same probability, we give some properties which we would desire this assignment of probabilities to possess. The "modified value" is then given by expression (1).

PROPERTY 1: An ordering and the reverse ordering should have the same probability.

Heuristically, Property 1 is desirable in that it is possible to bring up a proposal for voting in two ways: the proposal itself, or the contradictory proposal. If one proposal gives rise to a certain ordering, the contradictory should give rise to the reverse ordering of the players. Thus both orderings should have the same probability.

Another important consequence of Property 1 is that the value obtained for a game, and that for the dual game (see [4] for dual games) will be the same. For voting games, this means, as pointed out in [1], that the "winning power index" and the "blocking power index" will be the same.

Consider, finally, the case of the "pure bargaining game" B_2, a two-person game with characteristic function

$$v(\{1\}) = v(\{2\}) = 0$$

$$v(\{1, 2\}) = 1.$$

For this game, it seems reasonable to expect a value $\phi_1 = \phi_2 = 1/2$. In fact, there should be no modification of the value for a two-person game: since our modification depends only on the fact that certain coalitions are more likely to form, such a modification can only occur when there are at least three players.

PROPERTY 2: The removal of a subset, S, of the players, should not affect the probabilities assigned to the relative orderings of the remaining set, $N - S$, of players.

Heuristically, this means that the degree of affinity between two players should not be affected by the presence or absence of a third (though we might imagine two good friends breaking their friendship because of a third person).

An important consequence of this property is that the addition of dummies will have no effect on the value of a game. Thus, we can treat an n-person game as an $(n+1)$-person game by the addition

of a dummy: Property 2 guarantees that the value to the n original players will be the same in both cases.

These, then, are the properties desired. We give below a randomization scheme over the orderings of $N=\{1, 2, \ldots, n\}$ which will have both of these properties, and which has, we believe, a plausible heuristic interpretation. We give first, however, a pair of geometric lemmas.

LEMMA 1: Let $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ be points in Euclidean n-dimensional space, such that

$$\sum_1^n x_i^2 = \sum_1^n y_i^2, \tag{2}$$

and suppose that $(z_1, \ldots, z_n)$ is such that

$$\sum_1^n (x_i - z_i)^2 < \sum_1^n (y_i - z_i)^2 \quad . \tag{3}$$

Then, for every $t>0$,

$$\sum_1^n (x_i - tz_i)^2 < \sum_1^n (y_i - tz_i)^2.$$

PROOF: Because of (2), x and y are at an equal distance from the origin. Thus the boundary of the set of z satisfying (3) is a hyperplane passing through the origin. We see that both the set defined by (3) and its complement are convex. Suppose, then, that for some scalar $t>0$, the point tz does not lie in this set.

If $t>1$, then both 0 and tz lie outside (3). By convexity, z should also lie outside (3). If $0<t<1$, then both 0 and z lie in the closure of (3). By convexity, tz must also lie in the closure. Thus, it must lie on the boundary of (3); but by linearity, this means z must be on the boundary of (3). *In either case we get a contradiction.*

LEMMA 2: Let S_n be the unit sphere of Euclidean $(x+1)$-space

$$x_1^2 + x_2^2 + \ldots + x_{n+1}^2 = 1,$$

and let S_{n-1} be the intersection of S_n with the hyperplane $x_{n+1}=C$ where C is some constant such that $|C|<1$. (Thus S_{n-1} is an $(n-1)$-sphere.) Let T_{n-1} be a (Lebesgue) measurable subset of S_{n-1}, and let T_n be that subset of S_n defined by:

$$T_n = \left\{ x \;\middle|\; \begin{array}{l} \text{for some } y \epsilon T_{n-1} \text{ and some} \\ t>0,\ x_i = ty_i \text{ for } i=1, \ldots, n \end{array} \right\}.$$

Then

$$\frac{\lambda_n(T_n)}{\lambda_n(S_n)} = \frac{\lambda_{n-1}(T_{n-1})}{\lambda_{n-1}(S_{n-1})},$$

where λ_n and λ_{n-1} are Lebesgue n-dimensional and $(n-1)$-dimensional measure, respectively.

PROOF: The truth of this assertion can best be seen if we write

$$\lambda_n(T_n) = \int_{T_n} d\lambda_n$$

$$= \int_{T_{n-1}} \left[\int_{-1}^{1} f(x_1, \ldots, x_n = 1)\, dx_{n+1} \right] d\lambda_{n-1}$$

and

$$\lambda_n(S_n) = \int_{S_{n-1}} \left[\int_{-1}^{1} f(x_1, \ldots, x_{n+1})\, dx_{n+1} \right] d\lambda_{n-1}.$$

Now, $f(x_1, \ldots, x_{n+1})$ is an area function which depends only on the two numbers $x_1^2 + \ldots + x_n^2$ and x_{n+1}. Since $x_1^2 + \ldots + x_n^2$ is constant throughout S_{n-1}, it follows that the inner integrand is independent of the first n variables. Thus the inner integral depends only on c and can be taken outside the other integral. We will thus have

$$\lambda_n(T_n) = k\lambda_{n-1}(T_{n-1})$$

$$\lambda_n(S_n) = k\lambda_{n-1}(S_{n-1}),$$

which proves the Lemma.

THEOREM 1: Let S_{n-1} and S_n be an $(n-1)$-sphere and an n-sphere, respectively, such that S_{n-1} is embedded in S_n. Let $x^1, x^2, \ldots, x^p$ be points in S_{n-1}, let T_n be the set of all z in S_n such that

$$d(z,x^1) < d(z,x^2) < \ldots < d(z,x^p) \tag{4}$$

(where d represents the usual Euclidean metric), and let $T_{n-1} = S_{n-1} \cap T_n$. Then

$$\frac{\lambda_{n-1}(T_{n-1})}{\lambda_{n-1}(S_{n-1})} = \frac{\lambda_n(T_n)}{\lambda_n(S_n)}.$$

PROOF: We may choose a system of coordinates such that S_n and S_{n-1} are as in Lemma 2. Since $x^1, \ldots, x^p$ are all in S_{n-1}, we will have

$$\sum_{i=1}^{n} (x_i^j)^2 = 1 - c^2 \tag{5}$$

and

$$x_{n+1}^j = c, \tag{6}$$

for all $j, \ldots, p$.

Condition (6) means that, whether a point z, satisfies the condition $d(z, x^j) < d(z, x^k)$ is equivalent to whether

$$d(\bar{z}, \bar{x}^j) < d(\bar{z}, \bar{x}^k),$$

where $\bar{z}$, $\bar{x}^j$ and $\bar{x}^k$ are the projections of z, x^j and x^k into the plane of the first n coordinates. Condition (5) means that we can apply Lemma 1 to the projections of the $\bar{x}^j$ into this plane.

Now let $z \epsilon T_{n-1}$, let $t < 0$ and let $y \epsilon S_n$ be such that $y_i = tz_i$ for $i = 1, \ldots, n$. Then the following inequalities are equivalent:

$$d(z, x^j) < d(z, x^k)$$

$$d(\bar{z}, \bar{x}^j) < d(\bar{z}, \bar{x}^k) \qquad \text{(by above)}$$

$$d(t\bar{z}, \bar{x}^j) < d(t\bar{z}, \bar{x}^k) \qquad \text{(by Lemma 1)}$$

$$d(\bar{y}, \bar{x}^j) < d(\bar{y}, \bar{x}^k) \qquad (\text{since } \bar{y} = t\bar{z})$$

$$d(y, x^j) < d(y, x^k) \qquad \text{(by above).}$$

We see thus that $z \epsilon T_{n-1}$ if and only if $y \epsilon T_n$. It follows that T_n and T_{n-1} are as in Lemma 2, and thus the conclusion of Lemma 2 holds. This proves the theorem.

This geometric framework allows us to construct a randomization scheme which will have Properties 1 and 2. In fact, let us suppose that the several players in an n-person game are each assigned a point, x^j, in some Euclidean space of high dimension. (Two points will be close together if there is a high affinity between the corresponding players). Now these points, if in general position, will determine an $(n-2)$-sphere, S. An arbitrary point $z \epsilon S$ determines an ordering of the players $1, \ldots, n$: namely, the order of increasing distance of the points $x^1, x^2, \ldots, x^n$ from z. (This is true unless there are ties, but the points z which give rise to ties form a set of measure zero). *Then, to each ordering of the players, we assign a probability which is proportional to the measure of the set of all* z *which determine this ordering.*

We give an example of this in Figure 1 where a circle (determined by three points 1, 2, 3) is split into six arcs, each determining an ordering of the three players.

It is easy to see that such a randomization scheme will have the two desired properties.

To prove Property 1, we note that the sets corresponding to two mutually reverse orderings of the players will be antipodal sets, but antipodal sets on a sphere have the same measure. Hence an ordering and its reverse must have the same probability.

To prove Property 2, we note that removing one of the players from the game will (in the nondegenerate case) cause us to replace the sphere, S, with a lower-dimensional sphere S'. Theorem 1 guarantees that the relative orderings of the remaining players will have the same probabilities in the reduced game. (In a degenerate case, it may be all n points lie on p-sphere, where $p < n-2$, and that removal of a point will not change this. Property 2 is then trivial to verify.) We note that, if the points $x^1, \ldots, x^n$ are the vertices of a regular n-simplex, all the orderings have the same probability, i.e., the usual Shapley value is obtained.

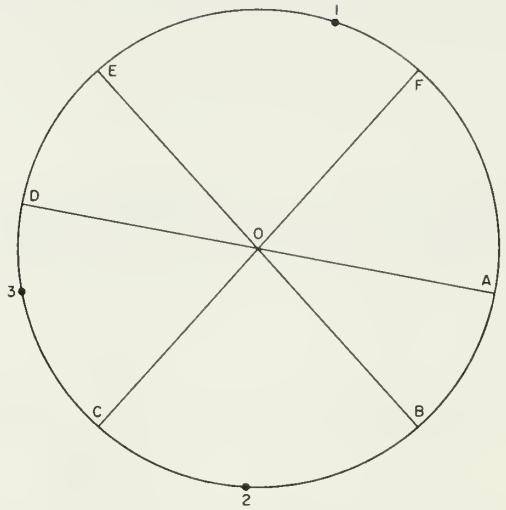


FIGURE 1

An interesting result now holds. Let us suppose that $n-1$ of the points x^j are kept fixed, while the remaining point, say x^n, is allowed to move over the sphere S. As x^n moves along S, then, so long as it avoids the other $n-1$ points, the $n-1$ hyperplanes given by

$$d(z, x^j) = d(z, x^n)$$

will move continuously through space. Thus the measures of the sets corresponding to each ordering will vary continuously with the position of x^n. It follows that the value to the players is a continuous function of the position of the point x^n except for discontinuities when x^n coincides with the other points. Now, let us assume x^n approaches another of the points, say x^{n-1}. If x^n and x^{n-1} are very close together, then for each $j=1, \ldots, n-2$ the two hyperplanes

$$d(z, x^j) = d(z, x^n)$$

and

$$d(z, x^j) = d(z, x^{n-1})$$

will be very close together. The angle between two such hyperplanes will be very small, and thus the orderings in which the two players n and $n-1$ do not appear consecutively will have very low probability. Thus, the value behaves as though the two players merged into a single player.

We give the several properties of this value in a summary theorem.

THEOREM 2: The scheme devised above gives a value which has Properties 1 and 2, and moreover, is individually rational, pareto-optimal, and continuous as a function of position, except for discontinuities where two or more points coincide.

The degenerate case, where all points lie on a sphere of lower dimension than $n+2$, is not without interest. In fact, a lower dimensional case is always easier to analyze. Now, there are conditions in

which we may assume that all the players can be represented, at least approximately, by points lying on a one- or two-sphere. This might be the case in politics, where a right-to-left spectrum is often posited (with no lack of political analysts to remind us that the extremes of right and left come together). Thus the parties in a legislature can be represented as points on the circumference of a circle (one-sphere) or, at worst, as on the surface of a two-sphere. We analyze such a legislature below.

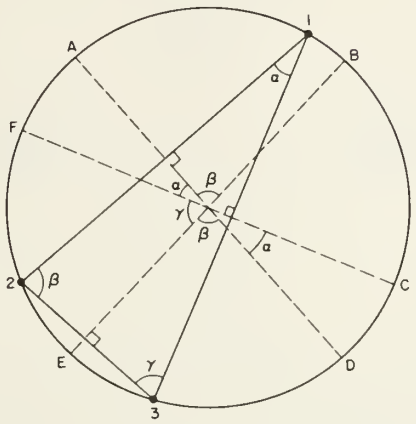


FIGURE 2

EXAMPLE 1: *The three-person games*

In a three-person game, we represent the players as the three vertices of a triangle with angles α, β, γ. (See Figure 2.) The triangle is inscribed in a circle with center 0.

The three perpendicular bisectors of the triangle cut the circle at six points A, B, C, D, E, F; it is then easy to see that the arcs determined will be:

$$FA = CD = \alpha$$

$$AB = DE = \beta$$

$$BC = EF = \gamma$$

This gives the ordering probabilities

$$\mathrm{P}(123) = \mathrm{P}(321) = \beta/2\pi$$

$$\mathrm{P}(132) = \mathrm{P}(231) = \gamma/2\pi$$

$$\mathrm{P}(312) = \mathrm{P}(213) = \alpha/2\pi$$

and thus we will have, as the modified Shapley value:

$$\phi_1 = \frac{1}{2\pi}\Big\{(\beta+\gamma)[v(\{1\}) + v(\{1,2,3\}) - v(\{2,3\})] + \alpha[v(\{1,2\}) + v(\{1,3\}) - v(\{2\}) - v(\{3\})]\Big\}.$$

with similar expressions for the other two players. In case v is the constant sum three-person game in (0,1) normalization, M_3, we will obtain the much simpler expression

$$\phi_1 = \alpha/\pi$$

$$\phi_2 = \beta/\pi$$

$$\phi_3 = \gamma/\pi$$

whereas, for the pure bargaining game B_3, also in (0,1) normalization, we have

$$\phi_1 = (\beta+\gamma)/2\pi$$

$$\phi_2 = (\alpha+\gamma)/2\pi$$

$$\phi_3 = (\alpha+\beta)/2\pi.$$

Generally speaking, if α is large, we find that point 1 is close to the middle of the line-segment $\overline{23}$. Thus, he can form a coalition easily with either 2 or 3, whereas the coalition $\{2, 3\}$ is hard to form. Then, for the majority game M_3, we find that player 1 has a considerable advantage. For the game B_3, however, he is at a disadvantage. This corresponds to the notion that, in a majority situation, the center party can often play one extreme against the other, but in a situation where all parties have a veto, it is usually the extreme parties that call the tune, with the center parties rarely determining anything.

If, on the other hand, α is small we find that player 1 is at a disadvantage in the game M_3, where the winning coalition $\{2, 3\}$ will form quite often, but is nearly as strong as the other two combined in the game B_3 (since, in fact, the other two will almost always act together).

EXAMPLE 2: *My Aunt and I*

This name was given by Maschler [2] to the five-person weighted majority game in which one player has three votes, and the other four players have one vote each. We will simplify it to a four-person game in which one player has two votes, and the others have one vote each.

The Shapley value for this game, according to the usual analysis, is

$$\phi_1 = 1/2$$

$$\phi_2 = \phi_3 = \phi_4 = 1/6$$

(where player 1 is the strong player). The game is, however, complicated by the fact that players 1 and 2 have a natural affinity for each other: in fact, player 1 is 2's aunt. Thus we may assume that the points 1 and 2 are closer together than the other points. We could assume that the four points determine a sphere. To simplify matters, however, let us assume that they all lie on a circle. Since the remaining players, 3 and 4 are strangers, we may assume that points 3 and 4 are equidistant from the midpoint, C, of the arc $\overset{\frown}{12}$. (See Figure 3.)

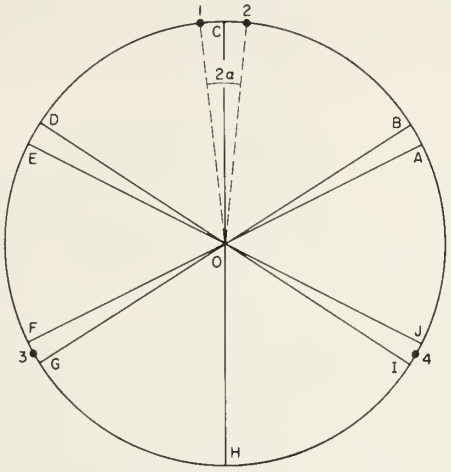


FIGURE 3

We draw, now, the perpendicular bisectors of the six sides of 1234. Two of these (namely, those of 12 and 34) coincide, so we obtain five bisectors, dividing the circle into 10 arcs $\widehat{AB}, \widehat{BC}, \ldots, \widehat{JA}$.

We know, now, that player 1 will be the pivotal player for any ordering in which he appears in either second or third position. Thus 1 is pivotal for any point on the arcs $\widehat{EF}, \widehat{FG}, \widehat{GH}, \widehat{JA}, \widehat{AB}, \widehat{BC}$. Now, the lines OC and OE bisect $\widehat{21}$ and $\widehat{13}$, respectively. We have, then

$$\widehat{CE} = \widehat{HJ} = 1/2\ \widehat{23}$$

so

$$\widehat{CE} + \widehat{HJ} = 2\pi/3 + \alpha$$

and so, remembering that 1 is pivotal except on arcs CE and HJ,

$$\phi_1 = 2/3 - \alpha/2\pi.$$

Similarly, we find that 2 is pivotal on the arcs $\widehat{CD}$ and $\widehat{HI}$. We have

$$\widehat{CD} + \widehat{HI} = \widehat{13} = 2\pi/3 - \alpha$$

and so

$$\phi_2 = 1/3 - \alpha/2\pi.$$

Similarly, we will have

$$\phi_3 = \alpha/\pi,\ \phi_4 = 0.$$

There is a lack of symmetry between players 3 and 4, due to the fact that 3 has been placed nearer to 1 (the strong player) than 4. Since this is in fact arbitrary, we might do better to replace these best values by $\phi_3 = \phi_4 = \alpha/2\pi$. This is not, however, very important. The important fact is that, as 1 and 2 approach each other, we will have

$$\lim_{1\to 2} \phi_1 = 2/3$$

$$\lim_{1\to 2} \phi_2 = 1/3$$

$$\lim_{1\to 2} \phi_3 = \lim_{1\to 2} \phi_4 = 0$$

and this limit is suggested as a "reasonable" value for the game "My Aunt and I." Similar analysis for the five-person game yields a value of $\phi_1 = 3/4$, $\phi_2 = 1/4$, $\phi_3 = \phi_4 = \phi_5 = 0$.

EXAMPLE 3: *The Knesset*

This research was begun while the author was at a conference in Jerusalem, Israel, October–November of 1965. During this conference, elections for the Knesset were held. A discussion with the Israeli members of the conference gave us a set of comparative positions for the several political parties (assumed to occupy approximately one half of a circle). These positions are given below in multiples of π; they vary from 0 to 1 since we assumed that only one half of the circle was occupied. We then treated this as a simple majority game, and obtained the modified values given in Table 1. In retrospect, it should be pointed out that a coalition government was eventually formed from which only the Rafi, Gahal, and Communist parties were excluded; the A. I., however, refused to join any coalition.

Heuristic. An interpretation of this model is perhaps necessary. The author feels that he should acknowledge a definite debt to the discussion in [1] of the Arrow paradox (pages 353–357).

TABLE I

Party	Ordinate	Seats	Value ϕ
New Communists	0.00	3	0.000
Communists	0.05	1	0.000
Poaelei Aguda	0.20	2	0.000
Agudat Israel	0.20	4	0.063
National Religious	0.20	11	0.063
Mapam	0.30	8	0.075
The Alignment (Mapai–Ahdut)	0.40	49	0.700
Haolan Haze	0.50	1	0.000
Independent Liberals	0.65	5	0.050
Rafi	0.80	10	0.050
Gahal	1.00	26	0.000

The political interpretation is paramount. Suppose, indeed, that the political parties in a country can be identified with definite positions on the "political spectrum." Suppose, then, that a political issue arises. Such an issue can generally also be identified with a point, P, on this spectrum. It seems reasonable that the jth party will favor the issue if the points x^j and P are close together, and oppose it otherwise. Thus, on each issue, a definite ordering of the players arises, in order of distance of the points x^j from the point P. Since we wish to know the probability that a given player will be the pivotal player, we look for the probability of obtaining a point which will give an ordering under which the given player is pivotal. But this is best done by measuring the set of such points and seeing what proportion it is of the set of all possible points. These sets are all unbounded, and thus will have infinite measure.

It is, however, easy to see that they are all cones with vertex at the center of the sphere determined by the x^j. Instead of measuring the cones, we measure their solid angles, or, equivalently, their intersection with the surface of this sphere. Thus our model is obtained.

BIBLIOGRAPHY

[1] Luce, R. D. and H. Raiffa, *Games and Decisions* (Wiley, New York, 1957).

[2] Màschler, M. and M. Davis, "The Kernel of a Cooperative Game," Nav. Res. Log. Quart. **12**, 223–259 (1965).

[3] Shapley, L. S., "A Value for n-Person Games," *Annals of Mathematics Studies*, *28*, Princeton, N.J., (1953), p. 307–318.

[4] Shapley, L. S., "Simple Games: An Outline of the Descriptive Theory," Behavioral Science **7**, 59–66 (1962).

[5] Shapley, L. S., "The 'Value of the Game' as a Tool in Theoretical Economics," RAND Corporation paper P–3658 (Santa Monica), (Aug. 1967).

AN APPLICATION OF LINEAR PROGRAMMING TO CONTINGENCY PLANNING: A TACTICAL AIRLIFT SYSTEM ANALYSIS

David C. Dellinger

Duke University

ABSTRACT

A linear programming application for the selection of aircraft for a tactical airlift fleet is described.

Linear programming has been used extensively over the past few years in the Department of of Defense to analyze strategic mobility problems [1]. These analyses seek a minimum cost combination of long-range aircraft, cargo ships, and prepositioned stocks; a combination which has the capability of meeting a broad array of contingency situations. In the linear programming formulation of the problem, the estimated demands for transportation necessary to deploy combat forces to each of the contingency situations are specified as constraints, and the numbers of aircraft, ships, etc., necessary to meet the demands are the variables. This paper describes an application of the same approach to a different mobility problem: the selection of aircraft for a tactical airlift fleet, the aircraft which provide mobility within the contingency area.

The characteristics of these two problems are common to many other problems and the approach described is applicable to a general class of problems. Essentially, the problems can be characterized as follows:

a. The decision variables are the numbers of different types of elements in a system.

b. The system is required to have the flexibility to perform a number of different sets of tasks.

c. Each of the possible elements in the system can perform a number of different tasks, but not necessarily with the same effectiveness.

The tactical airlift problem, has these characteristics. The elements in the system are the different types of candidate aircraft, and the problem is to decide on the number of each type to include in the tactical airlift fleet. The system composed of these aircraft must have the flexibility to perform a wide variety of missions or tasks. While all the tasks are similar in that they are to provide transport from point to point, they are different in many respects, e.g., distance traveled, type of airfield at origin and destination, type of cargo, etc. Moreover, the mixture of tasks differs from day to day. For example, on one day the tactical airlift system may be required to move a large amount of dense cargo over a long distance, while on the next day, it may be required to make many deliveries of small amounts to many different points. In regard to the third characteristic, each of the candidate aircraft is capable of performing a number of these tasks. For example, helicopters, which are generally recognized as being efficient in moving small loads over short distances, could also be used to move large amounts of cargo over long distances if the cargo were broken into small loads and refueling stops made en route. Conversely, large aircraft, which are generally recognized as being efficient in moving large loads over long distances, could also be used to move small loads over short distances.

Examples of nonmilitary problems which have these same characteristics include the problems of determining the mix of machines in a job shop, the mix of vehicles owned by a rental agency, the mixture of fire fighting vehicles in a fire department, the mixture of components in a health care delivery system, or the composition of an academic faculty capable of serving various mixes of students.

Designers have attempted to solve the tactical airlift problem with two divergent types of aircraft: those which are highly flexible and those which are highly specialized. For example, highly flexible aircraft capable of performing all the tasks required of the tactical airlift system have been developed. The V/STOL (Vertical or Short Takeoff and Landing) aircraft is such an aircraft. It combines the features of conventional aircraft and the helicopter and can duplicate the performance of either. On the other hand, highly specialized aircraft have been developed to perform specific tasks efficiently. Small conventional aircraft, for example, can deliver small loads over short distances between short runways very efficiently. A fleet composed entirely of either of these divergent types of aircraft may or may not constitute an efficient system.

To identify an efficient system for the tactical airlift problem requires an approach which:

a. Avoids suboptimization [2], i.e., does not require that the tasks to be performed by the tactical airlift fleet be divided into sub-groups for individual analysis.

b. Evaluates system or fleet flexibility rather than individual candidate aircraft flexibility.

c. Considers simultaneously all the capabilities of the candidate aircraft (since an aircraft in the system will normally have to perform a variety of tasks).

d. Realistically reflects the total set of relevant capabilities of the candidate aircraft.

One of the most common forms of analysis of the tactical airlift problem is a task-by-task cost-effectiveness analysis. In this form of analysis, candidate aircraft are compared on the basis of their efficiency in performing a single specified task. For example, a number of different aircraft may be compared on the basis of the efficiency with which they deliver a specified number of troops over a fixed distance to a point having a very small landing field. The aircraft which can perform this task at least cost is designated "best" for this particular task. The analysis is usually extended to include a variation in one or more of the parameters, e.g., distance, to find the points where the "best" aircraft changes. Such an evaluation tends to encourage, rather than avoid suboptimizations, i.e., it tends to encourage division of the total set of tasks into subsets and the selection of the "best" aircraft for each subset. Consideration of flexibility is restricted to single aircraft flexibility, and consideration of capability is restricted to the specified task. This form of analysis does permit a realistic representation of the capabilities of the candidate aircraft for the single task since both the task and the aircraft productivity can be described in great detail.

The linear programming formulation of the problem which follows avoids suboptimization by treating the problem as a system problem. The desired capability and flexibility of the entire system is specified and an entire system which meets the specification is sought as a solution to the problem. The desired flexibility of the system is specified in quantitative form and any candidate aircraft, whether flexible or highly specialized, can compete for a place in the system. Each candidate aircraft is evaluated on the basis of its capability to perform all the tasks specified for the system. Because the model is linear, however, it presents some problems in realistically reflecting the capabilities of candidate aircraft. Ways of dealing with these problems are described in the application section of this paper.

This model is also designed to address the problem of timing the replacement of aircraft already in the fleet by newer and more efficient aircraft, the so-called "fleet sequencing" problem. This is accomplished by permitting period by period decisions on the composition of the fleet over a series of

time periods of a fixed length. At each decision point, the option to continue to keep an old aircraft in the fleet or to replace its capability with a new aircraft is open. The linear programming solution will contain the sequence of fleets which minimizes the total discounted cost for all the time periods.

In the following sections a formal description of the model is given, the application of the model to the tactical airlift problem is described, the flexibility of the fleet is discussed, and a numerical example is given.

THE MODEL

The model is a standard linear program, see [3], consisting of an objective function to be minimized and a set of constraint equations (or inequalities). The unusual feature of the model is that it permits an explicit specification of the flexibility to be possessed by the fleet. This is accomplished by specifying a number of mission sets, each of which must be within the capability of any fleet in the set of feasible fleets. An optimal fleet is a feasible fleet which costs no more than any other feasible fleet. Note that the solution is not necessarily the least cost fleet capable of accomplishing *just* one of the mission sets, but each of the mission sets.

The constraint equations can be divided into four classes:

a. the requirements equations which assure that any feasible fleet will have the capability to accomplish each of the specified mission sets;

b. the fleet specification equations which define fleets have the specified degree of flexibility;

c. the fleet sequencing equations which permit aircraft from the fleet in one period to be included in the fleet for the following period; and,

d. a set of convenience equations which sum sets of variables to facilitate interpretation of the solution.

Each of these classes is discussed below using the following subscripts for the variables and parameters of the model:

i = type of aircraft	$i = 1$ to I,
j = operating base to which aircraft is assigned	$j = 1$ to J,
k = time period	$k = 1$ to K,
l = mission set	$l = 1$ to L,
m = mission type	$m = 1$ to M.

1. *Requirements Equations.* Let—

p_{ijklm} = productivity of aircraft type i, operating from base j in period k on mission m of requirement set l. The units of p_{ijklm} will be the same as the units of mission m, e.g., tons per day per aircraft;

w_{ijklm} = number of aircraft of type i, operating from base j in period k assigned to mission m of requirement set l; and

r_{klm} = number of units of mission m required in period k, mission set l.

The requirements equations are of the form,

$$\sum_{i=1}^{I} \sum_{j=1}^{J} p_{ijklm} w_{ijklm} \geq r_{klm}, \tag{1}$$

for $k = 1$ to K, $l = 1$ to L, and $m = 1$ to M.

Both the p_{ijklm} and r_{klm} are specified, and the w_{ijklm} are found in the solution to the linear programming

problem. The w_{ijklm} can be thought of as utilization variables since they show how the fleet determined in the solution to the linear programming problem can be utilized to accomplish the m simultaneous missions in mission set l of period k.

In matrix notation, the set of requirements equations can be represented as a diagonal matrix and two column vectors,

$$\begin{bmatrix} P_{11} & 0 & & & & 0 \\ \cdot & P_{12} & & & & \cdot \\ \vdots & 0 & \cdot & & & \vdots \\ \cdot & \cdot & & P_{kl} & & \cdot \\ \vdots & \vdots & & & \cdot & \vdots \\ 0 & 0 & & & & P_{KL} \end{bmatrix} \begin{bmatrix} W_{11} \\ W_{12} \\ \vdots \\ W_{kl} \\ \vdots \\ W_{KL} \end{bmatrix} \geqslant \begin{bmatrix} R_{11} \\ R_{12} \\ \vdots \\ R_{kl} \\ \vdots \\ R_{KL} \end{bmatrix}$$

in which P_{kl} is the diagonal matrix,

$$P_{kl} = \begin{bmatrix} p_{kl1} & 0 & 0 & \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & p_{kl2} & 0 & \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & p_{kl3} & \cdot & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & & & & & & \\ \cdot & \cdot & & p_{klm} & & & & \cdot \\ \cdot & \cdot & & & \cdot & & & \cdot \\ \cdot & \cdot & & & & \cdot & & \cdot \\ \cdot & \cdot & & & & & \cdot & \cdot \\ 0 & 0 & 0 & \cdot & \cdot & \cdot & \cdot & p_{klM} \end{bmatrix}$$

made up of the row vectors:

$$p_{kl1} = [p_{11kl1}, p_{12kl1}, \ldots, p_{1Jkl1}, p_{21kl1}, \ldots, p_{IJkl1}]$$

$$p_{kl2} = [p_{11kl2}, p_{12kl2}, \ldots, p_{1Jkl2}, p_{21kl2}, \ldots, p_{IJkl2}]$$

$$\vdots \qquad\qquad \vdots$$

$$p_{klm} = [p_{11klm}, p_{12klm}, \ldots, p_{1Jklm}, p_{21klm}, \ldots, p_{IJklm}]$$

$$\vdots \qquad\qquad \vdots$$

$$p_{klM} = [p_{11klM}, p_{12klM}, \ldots, p_{1JklM}, p_{21klM}, \ldots, p_{IJklM}],$$

$$W_{kl}=\begin{bmatrix} w_{11kl1} \\ w_{12kl1} \\ \vdots \\ w_{IJkl1} \\ w_{11kl2} \\ \vdots \\ w_{IJkl2} \\ \vdots \\ w_{11klM} \\ \vdots \\ w_{IJklM} \end{bmatrix} \quad \text{and} \quad R_{kl}=\begin{bmatrix} r_{kl1} \\ r_{kl2} \\ \vdots \\ r_{klM} \end{bmatrix}$$

P_{kl} is a matrix of aircraft productivities for mission set l in time period, k. It has M rows and $(I\times J\times M)$ columns. Normally, the P_{kl} are identical matrixes for all k and l.

W_{kl} is a column vector of aircraft assignments to the M mission in mission set l in time period k. It has $(I\times J\times M)$ elements.

R_{kl} is a column vector of requirements in mission set l in time period k. It has M elements.

2. *Fleet Specification Equations.* Let—

x_{Nik} = number of new aircraft of type i procured and operated in period k,

x_{Oik} = number of old aircraft, i.e., aircraft procured in the previous period, of type i operated in period k,

x_{Rik} = number of renovated aircraft of type i operated in period k. (Renovated aircraft are those procured in period k—2, operated as old aircraft in period k—1, and renovated at an additional cost to be operated in period k.)

The sum of the aircraft of each type included in the fleet for any period must be greater than or equal to the number required for any mission set in the period, i.e.,

$$\sum_{j=1}^{J}\sum_{m=1}^{M} w_{ijk1m} \leq x_{Nik}+x_{Oik}+x_{Rik}$$

$$\sum_{j=1}^{J}\sum_{m=1}^{M} w_{ijk2m} \leq x_{Nik}+x_{Oik}+x_{Rik} \tag{2}$$

$$\vdots$$

$$\sum_{j=1}^{J}\sum_{m=1}^{M} w_{ijkLm} \leq x_{Nik}+x_{Oik}+x_{Rik}$$

for $i=1$ to I and $k=1$ to K.

In matrix notation, the equations can be represented by

$$B\begin{bmatrix} W_{11} \\ W_{12} \\ \vdots \\ W_{kl} \\ \vdots \\ W_{KL} \end{bmatrix} \leq AX \quad \text{or} \quad -B\begin{bmatrix} W_{11} \\ W_{12} \\ \vdots \\ W_{kl} \\ \vdots \\ W_{KL} \end{bmatrix} + AX \geq 0,$$

where:

B is a $(K\times L\times M)$ by $(I\times J\times K\times L\times M)$ matrix of zeros and ones arranged to perform the summations indicated by the left side of the above Fleet Specifications, Eqs. (2) when used as a premultiplier,

A is a $(I\times K\times L)$ by $(3\times I\times K)$ matrix of zeros and ones arranged to perform the summations indicated by the right side of the Fleet Specification Eqs. (2) when used as a premultiplier of X, and

$$X^T=[x_{N11}, x_{O11}, x_{R11}, \ldots, x_{Nik}, x_{Oik}, x_{Rik}, \ldots, x_{NIK}, x_{OIK}, x_{RIK}] .$$

Note that it is this set of inequalities which enables the model to find a single minimum cost fleet for each period which is capable of accomplishing each of the mission sets specified for that time period. The fleet in the solution is not the sum of the aircraft necessary to meet the L mission sets, but it is the greatest of those required for all mission sets. This form has the effect of selecting aircraft on the basis of their productivities for each and every mission type.

3. *Fleet Sequencing Equations.* In order to relate the fleets for the K time periods, and, in particular, to assure that the use of old and renovated aircraft in one period is limited to the number of the same type of aircraft available from the previous period, the following constraint equations are necessary:

(3) $$X_{Ni(k-1)}\geq X_{Oik} \quad \text{or} \quad X_{Ni(k-1)}-X_{Oik}\geq 0,$$

and

(4) $$X_{Oi(k-1)}\geq X_{Rik} \quad \text{or} \quad X_{Oi(k-1)}-X_{Rik}\geq 0,$$

for $i=1$ to I and $k=1$ to K.

Equation (3) limits the number of old aircraft of the ith type in period k to the number of the same type of aircraft which were new in period $k-1$. Equation (4) limits the number of renovated aircraft of type i in period k to the number of old aircraft of the same type in period $k-1$. This is only one of many possible forms for these equations, of course, and it is based on the assumption that an aircraft can be included in the fleet for three time periods: in the period in which it is procured as a

new aircraft, in the succeeding period as an old aircraft, and in the third period after it is purchased as a renovated aircraft. The cost of including the aircraft in any fleet depends on its status; new, old, or renovated.

In matrix notation, this set of inequalities can be represented by

$$DX \geq 0.$$

where D is a $(2 \times I \times K) \times (3 \times I \times K)$ matrix of zeros and ones arranged to perform the summations indicated by Equations (3) and (4). D is of the form

$$D = \begin{bmatrix}
1 & 0 & 0 & 0 & -1 & 0 & & & & & & & & & & & & & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & -1 & 0 & & & & & & & & & & & & & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & & & & & & & & & & & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & & & & & & & & & & 0 \\
 & & & & & & & & & & & & & & & & & & & \\
. & & & & & & & & & & & & & & & & & & & . \\
. & & & & & & & & & & & & & & & & & & & . \\
. & & & & & & & & & & & & & & & & & & & . \\
 & & & & & & & & & & & & & & & & & & & \\
0 & & & & & & & & & & & & & & & & 0 & 1 & 0 & 0 & 0 & -1 & 0 \\
0 & 0 & & & & & & & & & & & & & & & 0 & 1 & 0 & 0 & 0 & -1
\end{bmatrix}.$$

4. *Objective Function.* Let—

c_{Nik} = cost of procuring and operating a new aircraft of type i in period k,

c_{Oik} = cost of operating an old aircraft of type i in period k, and

c_{Rik} = cost of renovating and operating an aircraft of type i in period k.

The objective function to be minimized is

$$\sum_{i=1}^{I} \sum_{k=1}^{K} c_{Nik}X_{Nik} + c_{Oik}X_{Oik} + c_{Rik}X_{Rik}.$$

In matrix notation, the objective function is CX, where $C = (c_{N11}, c_{O11}, c_{R11}, c_{N12}, \ldots, c_{N1K}, c_{O1K}, c_{R1K})$, and X is as defined earlier.

5. *The Complete Model.** The linear programming problem, then, is to find a nonnegative vector,

$$(W_{11}, W_{12}, \ldots, W_{kl}, \ldots, W_{KL}, X)^T,$$

$$+ b_{ijklm}\, w_{hklm} + \ldots \geq x_{Nik} + x_{Oik} + x_{Rik}$$

and

*The form of the constraint set suggests that the problem might best be solved by a decomposition technique, such as that described in [3]. The size of the problem in applications thus far has been on order of 1,000 to 1,500 constraints and it has not been necessary to resort to such techniques. Solutions have been obtained using IBM's Mathematical Programming System 1360 in 1–3 minutes.

to minimize CX subject to the following constraints:

$$\left[\begin{array}{cccc|c} P_{11} & 0 & \cdots & 0 & \\ 0 & P_{12} & & \vdots & 0 \\ \vdots & & \ddots & & \\ & & P_{kl} & & \\ & & & 0 & \\ 0 & & \cdots & P_{KL} & \\ \hline & -B & & & A \\ \hline & 0 & & & D \end{array}\right] \left[\begin{array}{c} W_{11} \\ W_{12} \\ \vdots \\ W_{kl} \\ \vdots \\ W_{KL} \\ X \end{array}\right] \geqslant \left[\begin{array}{c} R_{11} \\ R_{12} \\ \vdots \\ R_{kl} \\ \vdots \\ R_{KL} \\ 0 \end{array}\right]$$

6. *Convenience Equations.* To facilitate reading the solution, a number of additional variables and equations are included in the model. These equations merely sum selected variables and in no way influence the solution. They are not shown in the above formulation, but are included in the model.

a. Aircraft Deployment Summary. In the process of solving the problem, the optimal deployment of aircraft to operating bases is determined for each time period. The variables Z_{ijk} = number of aircraft of type i deployed to base j during period k, are determined by

$$Z_{ijk} \geqslant \sum_{m=1}^{M} w_{ijk1m}$$

$$Z_{ijk} \geqslant \sum_{m=1}^{M} w_{ijk2m}$$

$$\vdots \qquad\qquad \vdots$$

$$Z_{ijk} \geqslant \sum_{m=1}^{M} w_{ijkLm}$$

for $i=1$ to I, $j=1$ to J, and $k=1$ to K.

b. Aircraft Employment Summary. To show how the aircraft are employed to meet the specific requirements in each mission set, the variables, Y_{ikl} = number of aircraft of type i employed in period k for mission set l are determined by the equalities

$$Y_{ikl} = \sum_{j=1}^{J} \sum_{m=1}^{M} w_{ijklm}$$

for $i=1$ to I, $k=1$ to K, and $l=1$ to L.

APPLICATION TO THE TACTICAL AIRLIFT PROBLEM

To apply this model to the tactical airlift problem, it is necessary to specify:

a. The M types of mission to be performed,

b. The L mission sets for each of the k time periods, R_{kl},

c. The I candidate aircraft to be considered and the J operating bases to be used,

d. The productivities, p_{ijklm}, of the candidate aircraft when performing the M types of missions, and

e. The costs, C, of the candidate aircraft.

The M types of missions characterize the overall tactical airlift mission (all the types of tasks the system must be capable of performing) and the L mission sets define the flexibility and capability desired in the system. The productivities of the candidate aircraft are dependent upon the missions types, the basing pattern, and the operating rules of the aircraft, as well as the performance of the aircraft. Probably the greatest difficulty in applying the model centers around the definition of an appropriate set of missions. This section is devoted, for the most part, to this problem. It also includes comments on other factors affecting the productivities of the aircraft, mission set definitions and costs.

Mission Definitions

A specific type of mission is defined by specifying a particular set of values for mission variables such as the type and quantity of load to be moved, the pickup and delivery points for the move, the time permitted for the move, the takeoff and landing conditions, the delivery conditions, the distance to the pickup point, etc. The problem is to select M types of missions from the infinite number of types which could be defined by taking all possible values for all mission variables. Since the purpose of the model is to compare candidate aircraft on the basis of their productivities, this selection should be made so that the M types of missions selected portray those characteristics of the overall tactical airlift mission which are most significant in determining the productivities of the candidate aircraft. Consequently, the selection of a particular set of mission types is not independent of the set of candidate aircraft to be considered, and specific rules for selecting mission types cannot be given.

To provide insight into this problem, however, the following discussion indicates the variables considered significant in defining mission types in previous applications of the model. It also indicates the basis for selecting particular values for these variables and how the mission types defined were structured into the model.

1. *Time.* One of the difficulties in using the model is attaining a realistic representation of the time variable in what is basically a static allocation model. This problem can be handled by specifying for each mission type a particular time interval and expressing both the productivities, p_{ijklm}, and the units required, r_{klm}, in terms of the specified time interval. Any time interval may be used, provided it is long enough to permit the mission to be accomplished. For example, if it takes 2 hours to load and fly to the destination, a time period of less than 2 hours is meaningless. A time period of 1 day is normally used.

If a short response time is a realistic requirement for certain classes of missions (emergency) evacuation of troops, for example) an appropriate response time would be used for the mission and those aircraft which could not meet the specified response time would be assigned a productivity of zero for that mission type.

2. *Fixed Load Sizes.* One of the functions of the tactical airlift fleet is to deliver small quantities of urgently needed cargo. Another is the emergency air resupply of small combat units. Both generate

a need for movement of small loads. In addition, it may be desirable for tactical reasons to make some movements in small loads even though consolidation would be more economical. There are similar reasons for requiring that some movements be made in large loads.

The large load requirement can be handled by assigning a productivity of zero to those aircraft which cannot accommodate the required load size. The small load size is more troublesome, however, since there is no logical reason for not allowing a large aircraft to carry a small load. It is only necessary to prohibit the large aircraft from moving more than one load at a time. This type of mission can be described adequately by using two elements in the requirements vector, R_{kl}, to describe the mission: one for the total number of units to be delivered and one for the number of loads required. Each aircraft is assigned two productivities for the mission: one specifying the number of units it can move during the mission time interval and the other for the number of loads it can move during the time period. For example,

$$\ldots + p_{11klm}\, w_{11klm} + p_{21klm}\, w_{21klm} + \ldots \geqslant r_{klm}$$

$$\ldots + p'_{11klm}\, w_{11klm} + p'_{21klm}\, w_{21klm} + \ldots \geqslant r'_{klm},$$

where p_{ijklm}, w_{ijklm}, and r_{kl1} are as defined in the previous section, and

p'_{11kl1} = number of loads aircraft type 1 from operating base 1 can deliver on mission m,
p'_{21kl1} = number of loads aircraft type 2 from operating base 1 can deliver on mission m, and
r'_{klm} = number of loads required to accomplish m properly.

Both constraints would operate (with only one normally constraining) so that enough aircraft would be allocated to the activity to assure that both aspects of the mission could be satisfied. Even if a large aircraft type, perhaps one which could satisfy r_{klm} in a single load, were used, enough aircraft would be required to assure that deliveries could be made in the specified number of loads. Mission type 14 in Table I of the example which follows illustrates this formulation.

3. *Distances.* Variations in the distance to be flown produce a nonlinear variation in the productivities of the aircraft. Very short distances greatly reduce the productivity of aircraft because a larger share of the flying time is consumed in the takeoff and landing phases of flight than would be the case with long distances. For example, if 30 minutes of each flight is devoted to takeoff and landing activities (including approach for landing and climb out) which do not move toward the destination, one-half the potential productivity would be lost on a 150-mile flight at 300 miles per hour (30 minutes of productive time and 30 minutes for takeoff and landing) but only 12-1/2 percent is lost on a 1,050 mile flight at the same speed. Similarly, long distances can reduce payloads, and consequently productivities, because of the need for additional fuel. To permit these nonlinear variations in productivities to be properly represented in the model, several mission types can be defined which differ only in the distance between origin and destination. For example, missions 3 and 4 defined in Table I differ only in the distance parameter.

4. *Length of Available Runways or Landing Areas.* One of the most important factors in the design of intra-theater airlift systems is the effect of available runways and landing areas. To be effective in the early stages of a war or in a fluid situation, the aircraft should be capable of operating off runways which are either available or which can be quickly (and easily) prepared. Aircraft can be designed to

minimize this problem by reducing or eliminating the need for runways. Short takeoff and landing (STOL), vertical takeoff and landing (VTOL), or a combination, V/STOL aircraft can be designed for this purpose. In addition, conventional aircraft can partially overcome this problem through air dropping cargo (although at a reduced productivity and at a higher cost because of breakage and the extensive preparation necessary for airdrop).

To permit the impact of limited, or nonexistence of, runways to be evaluated, separate mission types are defined for different runway lengths. Aircraft which cannot operate off runways of the specified lengths are assigned productivities of zero for such missions. Aircraft, which normally require longer runways, but can operate off the shorter runways at reduced payloads, or those which can make airdrops, are also assigned reduced productivities for such missions.

5. *Load Types.* The productivities of the candidate aircraft are also affected by the types of loads carried. If the cargo has a high density, payload weights are generally greater than if the cargo has a low density. If the load is made up of passengers, the payload measured in weight may be very low even though the aircraft is fully utilizing its passenger carrying capability. For these reasons, cargo missions are subdivided by type of cargo, and separate mission types are defined for passenger movements. Table I shows several examples of mission classifications based on load types.

Productivity Variations

Productivities of the candidate aircraft can also be affected by factors which are independent of the mission types. Two such factors, the location of the operating bases for the aircraft and the use of combinations of aircraft for the same mission, have been structured into the model.

1. *Location of Operating Bases.* Candidate aircraft are identified both by aircraft type and by the location of its operating base. Productivity calculations include the nonproductive time required to fly the empty aircraft to the pickup point and to return the empty aircraft to its operating base upon completion of its mission. Obviously, those aircraft which are closest to the pickup and delivery points for any one mission will have the highest productivities for that mission.

2. *Combinations of Aircraft.* It is sometimes advantageous to use two different aircraft in combination for certain types of missions. For example, conventional aircraft and helicopters may be used together to perform missions where the distance is too great for the helicopters, and the landing field at the destination is unsuitable for conventional aircraft, missions which neither could perform independently. The conventional aircraft can move the cargo, or passengers, to a suitable landing field close to the destination and helicopters can move the load to its final destination. This operation, usually called a tandem operation, provides an alternative to be compared with long-range Vertical-Takeoff-and-Landing aircraft. The following illustrates how a tandem operation using aircraft type i from operating base j and aircraft type i' from operating base j' can be structured into the model.

A term of the form,

$$\ldots + p_{hklm}\, w_{hklm} + \ldots \geq r_{klm},$$

in which

p_{hklm} = productivity of one unit of tandem operation h on mission m of mission set l of time period k, and

w_{hklm} = number of units of tandem operation h used on mission m of mission set l of time period k, replaces a standard term in the Requirements Equation (1). In addition, to account for the aircraft used, terms of the form,

$$+ b_{i'j'klm}\, w_{hklm} + \ldots \geqslant x_{Nik} + x_{Oik} + x_{Rik}$$

in which

b_{ijklm} = number of aircraft of type i operating from base j required for each unit of tandem operation h, are added to the Fleet Specification Equations (2). (A corresponding modification is necessary in the matrix B.)

Mission Sets

With an appropriate set of mission types and corresponding productivities defined, the next step in applying the model is to specify mission sets, R_{kl}, which appropriately describe the desired range of flexibility desired in the fleet. A mission set is a vector whose mth element is the number of units of mission type m in the set. Each mission set is a combination of missions which any feasible fleet must be capable of performing simultaneously. It is desirable to select as small a number of mission sets as possible to define the desired system flexibility.

In practice, a mission set is usually considered to be those missions which must be accomplished on a particular day. The sets chosen for the model are those which present the greatest variation on an element by element basis. For example, during the early phases of a war, one might expect the number of missions related to deploying combat forces within the theater to be high while the number of missions related to delivering supplies to the few forces in the theater to be low. As the deployment of combat forces is completed, the reverse would probably be true. During a counter-attack or an emergency withdrawal of forces different mission sets would be expected.

Costs

The costs which are of primary interest to force structure analysts are the costs of maintaining during peacetime an airlift system which has the capability of performing a given set of missions during wartime. The cost of operating an airlift fleet during wartime is of secondary importance for at least two reasons. One is that, once a decision is made to maintain an airlift fleet, the peacetime costs associated with the fleet are certain while any wartime costs are problematical. In fact, one of the purposes of maintaining military forces is to deter* the war, and if military planners are successful, wartime costs would not be incurred at all. The other reason is that the incremental cost of switching from peacetime to wartime operations is small relative to the peacetime procurement and operating cost. In order to maintain a tactical airlift fleet in an appropriate state of readiness, it is necessary to operate fully manned units during peacetime, albeit at a reduced level. Moreover, when future costs are discounted to the present as they are in this model, the differences between peacetime costs, particularly procurement cost which occurs early, and wartime costs, which could only occur sometime after procurement of the aircraft, are even more pronounced.

In application of this model, only the peacetime costs of maintaining the fleet in a state of readiness to deploy and operate in the event of a war have been considered. The costs of actually operating during a war have not been considered. So the model is designed to find a tactical airlift fleet which has the capability defined by the specified mission sets and which imposes the least peacetime cost on the government to procure and maintain in a state of readiness. These costs are simply the incremental costs of adding an aircraft to the peacetime fleet. They include the costs of adding an aircraft (or a squad-

*Clifford, Clark M., "Statement of Secretary of Defense Clark M. Clifford, The Fiscal Year 1970–74 Defense Program and 1970 Defense Budget," p. 68.

ron of aircraft) to the fleet, the cost of necessary training and support facilities, and the cost of operating the aircraft in peacetime as necessary to maintain the desired state of readiness.

Since the model can only handle linear cost functions, fixed costs such as the cost of research and development for new aircraft are not included in the cost function, but must be added manually to the total cost determined by the model. To assure that a minimum cost fleet is actually found when the fixed costs are included, it is necessary to compare the cost of all possible fleets after including the fixed cost. Because the number of candidate aircraft in actual application has been small, it has been feasible to find minimum cost fleets composed of all possible combinations of types of aircraft and make such comparisons after adding fixed costs.

All costs are discounted to a present value at an appropriate interest rate. Prior to discounting, estimated cash flows are converted to a series of year-end payments which approximate the expected series of payments which would result from adding aircraft to the fleet. This series of payments, discounted to its present value, is the cost coefficient for aircraft in each period. If a new aircraft is procured during the time period, l, the entire procurement cost is charged to that period, even though the aircraft could be operated in subsequent time periods. If it is operated in subsequent time periods, only the operating costs are charged during those periods. The size of the fleet actually procured is larger than the number given by the linear programming solution by an amount equal to the number of aircraft expected to be in major overhaul (and therefore not available for missions) and the number expected to be lost through attrition. The cost of the additional aircraft procured for these purposes is proportional to the size of the required operational fleet and is included in the cost coefficients of the individual aircraft in the model.

Fleet Flexibility

It is important to note that any fleet which satisfles the requirements equations not only meet the L specified mission sets, but any convex combination of the mission sets as well. For each time period, k, with L mission sets, the following must hold:

$$
\begin{array}{lllll}
P_{k1}W_{k1} & & & & \geq R_{k1} \\
 & P_{k2}W_{k2} & & & \geq R_{k2} \\
 & & \cdot & & \vdots \\
 & & & \cdot & \vdots \\
 & & & & P_{kL}W_{kL} \geq R_{kL}
\end{array}
\tag{5}
$$

and

$$
\begin{array}{lllll}
B_{k1}W_{k1} & & & & \leq A_{k1}X_k \\
 & B_{k2}W_{k2} & & & \leq A_{k2}X_k \\
 & & \cdot & & \vdots \\
 & & & \cdot & \vdots \\
 & & & & B_{kL}W_{kL} \leq A_{kl}X_k,
\end{array}
\tag{6}
$$

where P_{kl}, R_{kl}, and W_{kl} are as defined earlier; X_k is vector made up of the elements in X pertaining to period k; the B_{kl} are the submatrices of matrix B; and the A_{kl} are the submatrices of matrix A pertaining to mission set l of period k. Recall that matrices B and A are made up of zeros and ones arranged to perform the following summations:

$$\sum_{j=1}^{J}\sum_{m=1}^{M} w_{ijk1m} \leq x_{Nik} + x_{Oik} + x_{Rik}$$

$$\sum_{j=1}^{J}\sum_{m=1}^{M} w_{ijk2m} \geq x_{Nik} + x_{Oik} + x_{Rik}$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$\sum_{j=1}^{J}\sum_{m=1}^{M} w_{ijkLm} \geq x_{Nik} + x_{Oik} + x_{Rik}$$

for $i=1$ to I and $k=1$ to K.

Since for any time period, k, the summations are identical for each mission set, l, both the B_{kl} and A_{kl} submatrices are identical for all l and can be written B_k and A_k. Moreover, if, as is the usual case, the productivity matrix, P_{kl}, is the same for all mission sets, l, in period, k,

$$P_{k1} = P_{k2} = \ldots = P_{kl} = \ldots = P_{kL},$$

and can be identified by P_k. Consequently, (6) can be rewritten:

$$\begin{aligned} B_k W_{k1} &\leq A_k X_k \\ B_k W_{k2} &\leq A_k X_k \\ &\vdots \\ B_k W_{kL} &\leq A_k X_k \end{aligned}$$

and it follows that

$$B_k(\alpha_1 W_{k1} + \alpha_2 W_{k2} + \ldots + \alpha_L W_{RL}) \leq A_k X_k, \tag{7}$$

if $0 \leq \alpha_l \leq 1$ and $\sum_{l=1}^{L} \alpha_l = 1$. In words, any convex combination of utilization, W_{kl}, can be obtained from X_k.

Furthermore, since $P_k = P_{kl}$ for all l, (5) can be rewritten:

$$\begin{aligned} P_k W_{k1} &\geq R_{k1} \\ P_k W_{k2} &\geq R_{k2} \\ &\vdots \\ P_k W_{kL} &\geq R_{kL} \end{aligned}$$

and

$$P_k(\alpha_1 W_{k1} + \alpha_2 W_{k2} + \ldots + \alpha_L W_{kL}) \geq \alpha_1 R_{k1} + \alpha_2 R_{k2} + \ldots + \alpha_l R_{kL} \tag{8}$$

if $0 \leq \alpha_l \leq 1$ and $\sum_{l=1}^{L} \alpha_l = 1$. Or in words, any convex combination of the mission sets, R_{kl}, for period, k, can be accomplished by the same convex combination of the utilizations, W_{kl}. Taking (7) and (8) together, we can conclude that any convex combination of the mission sets, R_{kl}, can be accomplished by the fleet, X_k, for that period. The significance of this point is that the solution specifies a fleet which is not only capable of satisfying the specific mission sets defined for the model, but the entire envelope of mission sets defined by the above convex combination. In general, the fleet, X_l, has the capability to meet any mission set which is in the convex combination of the mission sets specified for the time period.

A NUMERICAL EXAMPLE*

In this section, a numerical example of the application of the model to a tactical airlift problem is presented, and the results are compared to those one might obtain using a mission by mission cost-effectiveness analysis of candidate aircraft. The cost of the fleet determined by system optimization using the linear programming model is compared with the cost of the fleet determined by suboptimization on a mission level (the latter process) to indicate the relative cost of such suboptimization.

TABLE I. *Mission Definitions*

Mission type		Limit on load size	One-way distance (miles)	Runway length [a]		Load type	Units of measure	Maximum time to complete mission (days)
Number (*m*)	Description			Takeoff (ft)	Landing (ft)			
1	Deploy forces	None	300	5,000	5,000	Equipment	Tons	1
2	Personnel movement	do	300	5,000	5,000	Troops	Troops	1
3	Transport supplies	do	75	5,000	5,000	Bulky cargo	Tons	1
4	do	do	200	5,000	5,000	do	do	1
5	do	do	200	5,000	5,000	Dense Cargo	do	1
6	do	do	75	5,000	0	Bulky Cargo	do	1
7	do	do	400	5,000	0	do	do	1
8	do	do	75	5,000	1,000	do	do	1
9	do	do	400	5,000	1,000	do	do	1
10	do	do	75	5,000	2,000	do	do	1
11	do	do	400	5,000	2,000	do	do	1
12	do	do	75	5,000	3,000	do	do	1
13	do	do	400	5,000	3,000	do	do	1
14	do	Maximum [b]	25	0	0	do	do	1
14a [c]							Sorties	1
15	do	None	50	5,000	0	do	Tons	1/2
16	do	do	400	5,000	0	do	do	1/2
17	Transport troops	do	75	5,000	5,000	Troops	Troops	1
18	do	do	200	5,000	5,000	do	do	1
19	do	do	400	5,000	5,000	do	do	1
20	Redeploy forces	do	50	0	0	Troops and equipment.	Tons	1/2
21	do	do	150	0	0	do	do	1/2

[a] Zero indicates that vertical takeoff, landing, or air drop is necessary.

[b] Maximum load size is one day's consumption by a battalion size unit.

[c] Mission type 14 requires two constraint equations and, consequently, two entries in each mission set. The notation system defined earlier does not provide for such double entries, but it can easily be modified to do so.

*This example was taken from a paper presented at the 15th International Meeting of The Institute of Management Sciences in September 1968, "Linear Programming in the Analysis of the Tactical Airlift Systems," by David C. Dellinger and E. George Riedel.

This simplified example considers three types of aircraft, three mission sets, 21 mission types, one operating base, and one time period. The upper limits on the indices are: $I=3$, $J=1$, $K=1$, $L=3$, and $M=21$. The data for the example are displayed in Tables I, II, and III. To facilitate relating the example for the general model, subscripts $j=1$ and $k=1$ are included in all variable notation.

Table I defines the characteristics of the 21 mission types. Table II displays the productivities and costs of the three types of aircraft considered as candidate aircraft. Aircraft 1 is a conventional aircraft which requires about 3,000 feet of runway for normal payloads, but which can operate off 2,000 feet at reduced loads. Aircraft 2 is a short-range vertical takeoff and landing aircraft in the helicopter class, and aircraft 3 is a long-range, VTOL aircraft capable of performing any of the mission types. Table III shows the three mission sets which any acceptable fleet must be capable of performing.

TABLE II. *Productivities* [a] *and Cost of Candidate Aircraft*

Mission type number (m)	Units [b] of measure	Productivities of candidate aircraft i			
		$i=1$ P_{111lm}	$i=2$ P_{211lm}	$i=3$ P_{311lm}	1 & 2 Tandem [d] P_{h1lm}
1	Tons	4000	0	18	0
2	Troops	117	0	103	0
3	Tons	42	8	32	0
4	do	19	0	17	0
5	do	37	0	35	0
6	do	0	12	52	0
7	do	0	0	23	18
8	do	0	12	49	0
9	do	0	0	18	0
10	do	5	12	49	0
11	do	3	0	18	0
12	do	64	12	49	0
13	do	14	0	18	0
14	do	0	41	108	0
14a	Sorties [c]	0	2.5	12	0
15	Tons	20	22	38	0
16	do	4	0	80	18
17	Troops	288	70	218	0
18	do	130	0	120	0
19	do	86	0	83	0
20	Tons	0	12	14	0
21	do	0	33	50	0
Cost	*Dollars (in millions) per aircraft*	5.180	5.028	19.688	(e)

[a] Productivities are the same for all values of l, i.e., all mission sets.

[b] Productivities are given in units delivered per aircraft during the time permitted to complete the mission.

[c] A sortie is a single delivery.

[d] A "tandem" operation makes deliveries in two legs using one aircraft type for one leg and the other aircraft type for the second leg. The productivities in this example are based on the simplifying assumption that exactly one of each type of aircraft is required to produce the indicated delivery rate. Any appropriate ratio of aircraft can be used.

[e] Aircratt used in tandem are costed as individual aircraft.

TABLE III. *Mission Sets* R_{11}
(Units required per mission set (r_{11lm}))

Mission type number (m)	Units	Mission set number (l)		
		1	2	3
1	Tons	4,000	0	3,000
2	Troops	3,000	0	2,000
3	Tons	100	500	500
4	do	300	400	400
5	do	600	60	60
6	do	900	1,200	1,200
7	do	250	600	600
8	do	150	150	150
9	do	54	54	54
10	do	60	60	60
11	do	250	250	250
12	do	450	600	600
13	do	150	300	300
14	do	1,000	0	800
14a	Sorties	60	0	150
15	Tons	20	40	1,000
16	do	10	15	800
17	Troops	0	2,000	2,000
18	do	0	900	900
19	do	0	4,000	4,000
20	Tons	0	1,500	0
21	do	0	6,000	0

TABLE IV. *Solution To Example Problem*
(Number of aircraft used to satisfy constraints (w_{i11lm}))

Mission type number	Mission set $l=1$ Aircraft type, i			Mission set $l=2$ Aircraft type, i			Mission set $l=3$ Aircraft type, i		
(m)	1	2	3	1	2	3	1	2	3
1	200						150		
2	26						17		
3	3			12				63	
4	16			21			21		
5	16			2					
6			17			23		100	
7	[a]5	[a]5	7	[a]33	[a]33		[a]11	[a]11	18
8		5	2			3		13	
9			3			3			3
10			1			1		5	
11			14			14			14
12	7			9			7	15	
13	10			22			22		
14		24						60	
14a		(24)						(60)	
15		1		2				45	
16			1	4					10
17				7				28	
18				7			7		
19				46			46		
20					125				
21					182				
Total used	283	34	45	165	340	44	283	340	45
Slack	0	306	0	118	0	[b](1)	0	0	0
Total in system	283	340	45	283	340	45	283	340	45
System cost $4,061.42 million									

[a] Tandem activity using both aircraft types 1 and 2 in combination.
[b] Rounding error.

The solution vector contains 252 possible elements* for the w_{ijklm} and 3 possible elements for x_{i11lm}.

The solution to the problem is a fleet consisting of

$$x_{N11} = 283 \text{ type 1 aircraft}$$
$$x_{N21} = 340 \text{ type 2 aircraft}$$
$$x_{N31} = 45 \text{ type 3 aircraft}$$

which costs \$4,061.42 million. Table IV shows the utilization variables, w_{i11lm}, rounded† to the nearest integer. The entries are the numbers of each type of aircraft utilized for each mission type in each mission set. Those used in tandem operations are indicated by footnote. The number of unused, or slack, aircraft are shown at the bottom of the chart.

TABLE V. *Independent Solutions for Each Mission Set*

(Mission-by-mission cost effectiveness analysis)

Mission type number (m)	Mission set l=1 Aircraft type, i			Mission set l=2 Aircraft type, i			Mission set l=3 Aircraft type, i		
	1	2	3	1	2	3	1	2	3
1	200						150		
2	26						17		
3	3			12			12		
4	16			21			21		
5	16			2			2		
6			17			23			23
7	[a]14	[a]14		[a]33	[a]33		[a]33	[a]33	
8			3			3			3
9			3			3			3
10			1			1			1
11			14			14			14
12	7			9			9		
13	10			22			22		
14		24							13
14a		(24)							(13)
15		1			2			46	
16			1			1			10
17				7			7		
18				7			7		
19				46			46		
20					125				
21					182				
Total used	292	39	39	159	342	45	325	79	67
Slack	0	0	0	0	0	0	0	0	0
Total	292	39	39	159	342	45	325	79	67
System cost	\$2,474 million			\$3,431 million			\$3,399 million		

[a] Tandem activity using both aircraft types 1 and 2 in combination.

*Including activities for which the productivities are zero. The number of possible elements is $(I \times J \times M) \times (K \times L) = (4 \times 21)(1 \times 3) = (84)(3) = 252$. Note that the tandem operation is considered a separate type of aircraft for this purpose.

†The solution shown will not provide the exact capability specified for each constraint because of rounding. For example, the 26 aircraft used to satisfy mission type 2 in set 1 are capable of delivering only 147 tons per day rather than the specified 150. The solution obtained from the linear program was noninteger and did precisely meet the constraints. The solution obtained by rounding is not necessarily an optimal integer solution.

results were obtained by solving the linear programming problem for one mission set at a time, but could be easily predicted by calculating for each aircraft type the cost per unit of productivity for each mission type and selecting the aircraft having the lowest cost per unit to satisfy that mission type. For example, the cost per unit of accomplishing mission type 1 is 5.180/20 = $.259 million per ton for aircraft type 1 and 19.688/18 = $1.094 million per ton for aircraft type 3. By this criterion, aircraft type 1 is "best" for that mission type. Similar calculations for mission type 10 shows that the unit costs are $10.36, $.419, and $.402 million per ton for aircraft types 1, 2, and 3, respectively. Type 3 would obviously be selected for that mission type. Where double constraints are necessary (mission type 14), the unit costs for the constraint which is constraining will determine the most economical candidate. For example, constraints 14 and 14a are used to describe a single mission type and the cost per unit of productivity are:

Constraint	Cost per unit of productivity Dollars (in millions) per unit			
	Aircraft type			Units
	1	2	3	
14.....	N/A	0.1226	.18229	Tons
14*a*.....	N/A	2.0112	1.64067	Sorties

So, if constraint 14 dominates, aircraft type 2 would be selected, but if constraint 14a dominates, aircraft type 3 would be selected. This explains why aircraft type 2 was used to satisfy constraints 14 and 14a in mission set 1 of Table V and aircraft type 3 was used for the same mission in mission set 3. (Because the same mix of aircraft must be capable of meeting any of the three mission sets, the entries in Table IV cannot be determined so easily.)

These two sets of solutions illustrate some interesting points regarding the use of this model for selecting aircraft for an intra-theater airlift system.

1. *Cost of Flexibility.* Adding flexibility to an airlift system will increase its cost. In this example, the system having enough flexibility to meet all three mission sets costs $662 million (4061 − 3399) more than the system capable of meeting only mission set 3. These figures are based on the least cost way of obtaining the flexibility for meeting any of the specified mission sets. An alternative way of obtaining the desired flexibility sometimes suggested (called the worst case approach) is to select a system composed of the greatest number of aircraft of each type required for any mission set. A system composed in this way for this example (see Table V) would contain 325 type 1 aircraft, 342 type 2 aircraft, and 67 type 3 aircraft. Clearly it would have the desired flexibility, but its cost would be $4,722 million, $661 million more than the least cost system having the desired capability.

2. *Form of Flexibility.* One way of obtaining the desired flexibility suggested earlier is to select enough of a single aircraft type which has the flexibility to perform any of the intra-theater missions. An alternative is to select a system composed of a number of different aircraft which together have the desired flexibility. This model provides a way to examine both alternatives simultaneously. Note in Table II that aircraft type 3 is capable of performing all the mission types listed, so a system composed entirely of aircraft type 3 is a possible alternative and was considered (but not selected) as a

For comparison, Table V shows the least cost solution which would be obtained if a separate fleet were designed for each of the mission sets using the "best" aircraft for each mission type. These possible solution to the linear program. A system composed entirely of type 3 aircraft which has the desired capability and flexibility would consist of approximately 450 type 3 aircraft and would cost $8,859.6 million. A system composed in this way would be economically competitive only if the cost per aircraft were reduced from $19.67 million to about $9 million.

3. *Basis for Selecting Aircraft.* With this model, aircraft are considered for selection on the basis of their capability to perform a wide range of missions, not just the mission for which they are best suited. It is possible that an aircraft which is superior to another for one type of mission may not be selected for that mission because it is inferior in the performance of other missions. This is illustrated in the above example. Note that aircraft type 3 is, on a dollar per ton of delivery capability basis, the most efficient way available of delivering cargo under conditions specified by mission type 6, i.e., 19.688/52 = 0.38 million dollars per ton of capability versus 5.028/12 = 0.42 million dollars per ton of capability for aircraft type 2. (Type 1 cannot perform the mission.) In Table IV, however, this mission is accomplished by aircraft type 2 in mission set 3 even though the cost per ton is greater. This is because the aircraft type 2 can also be used to perform either of the short distance redeployment of forces missions specified by mission types 20 and 21. When a single aircraft, either 2 or 3, is considered to perform combinations of missions 6, 20, and 21, aircraft type 2 is less costly. This is illustrated in Table VI where we examine the capability obtainable for these three missions by expending $1 million on either aircraft types 2 or 3.

TABLE VI. *Capability Obtainable From Dollars (in million) Investment in Aircraft Type 2 or 3*

Units of productivity

Mission type number	Aircraft type	
	2	3
6	[a] 2.22	2.63
20	2.22	0.71
21	6.50	0.25

[a] Productivity divided by cost.

As the previous calculations showed, more capability per dollar can be obtained for mission type 6 alone by procuring aircraft type 3 than by procuring aircraft type 2. However, since the same aircraft can be used either for pairs of missions 6 and 20 or for pairs of missions 6 and 21 (20 and 21 must be met simultaneously) we find that aircraft type 2 provides the greater combined capability. When used for missions 6 and 20, aircraft type 2 provides 4.44 units of productivity (2.22 for mission 6 and 2.22 for mission 20), while aircraft type 3 provides only 3.34 (2.63 + 0.71). When used for missions 6 and 21, aircraft type 2 provides 8.72 (2.22 + 6.50) units of productivity while aircraft type 3 provides only 2.88 (2.63 + 0.25). The implication is that aircraft optimally designed for a particular mission may not be optimal for the airlift system. As a corollary, when considered in a system context, it may be efficient to procure aircraft to be used on missions for which they are inferior on mission by mission cost-effectiveness basis. This is illustrated by the fact that aircraft type 2 is used in six mission types (mission types 3, 6, 8, 10, 12, and 17) for which it is *not* most efficient on a mission-by-mission analysis. The point is, of course, that the objective is to select aircraft for an efficient mix—not to use aircraft in the missions for which they are most efficient.

CONCLUDING COMMENTS

This model, with slight variations, has been used extensively in the Office of the Assistant Secretary of Defense (Systems Analyses), in Headquarters USAF, and in the USAF Tactical Air Command in their analyses of the tactical airlift problem, and its use has contributed to a better understanding of the tactical airlift problem. It has provided a means of integrating inputs from two basic areas of knowledge essential for proper analysis of the problem—aircraft technology and military applications—and of synthesizing efficient airlift fleets. From aircraft technology come estimates of what is possible in the way of candidate aircraft and estimates of their performance and cost. From military applications come estimates of the number and types of missions to be performed. The model integrates these two inputs, and, for a given combination of mission sets and candidate aircraft, can identify an efficient fleet.

While this model has been used to solve specific problems in the analysis of force structure decisions, its greatest value has probably been in (1) providing a unifying structure for addressing a broad range of related problems, (2) focusing attention on the significant issues, (3) guiding the analysis of missions to be performed, and (4) suggesting efficient innovations in aircraft technology.

For years, the debates which surround the selection of tactical airlift aircraft (and other military systems as well) have centered on aircraft characteristics, rather than on the tasks or missions to be performed. Characteristics such as speed, payload, cargo compartment size, and airfield landing and takeoff requirements, have been debated and specified independently, or at best, on the basis of a single pure mission analysis. This model focuses attention on the task to be accomplished, and permits the aircraft to be evaluated on the basis of their productivity, not on their characteristics. If the high speed of an aircraft, for example, does not make it a more efficient competitor, it will not be selected for membership in an efficient fleet. In economic terminology, the model forces one to look at outputs, not the characteristics of the production process.

The model is, of course, a requirements model, i.e., the solution is determined by the specified mission sets, and unfortunately, military planners can at best obtain only rough forecasts of future demands for tactical airlift. However, rather than being a "slave" to specific requirements, the model can be used to guide the military planners. An investigation of the values of dual variables, for example, can identify the mission sets which are the most costly to meet. A careful analysis of these missions may lead to their reduction or elimination or a search for alternative (non-air) ways of meeting these basic transportation needs. The model also provides a means for the economic evaluation of innovative mission types, i.e., new ways to use airlift to accomplish jobs more efficiently. In general, it can be used to generate economic data to support the analysis of the tactical airlift mission.

On the other input side, aircraft technology, the model can be used to do more than evaluate candidate aircraft suggested by aircraft designers, although this in itself is a significant contribution. It can be used to suggest technological innovations which would increase the efficiency of the fleet. For example, an investigation of the characteristics of aircraft which compete well in the model, should suggest a design which would be even more efficient.

It is in the role of providing a precise structure and conceptual framework for the analysis and synthesis of tactical airlift systems and a basis for communication between the military planners and aircraft designers that this model can be most useful.

REFERENCES

[1] Fitzpatrick, G. R., J. Bracken, M. J. O'Brien, L. G. Wentling, and J. C. Whiton, "Programming the Procurement of Airlift and Sealift Forces: A Linear Programming Model for Analysis of the Least-Cost Mix of Strategic Deployment Systems," Nav. Res. Log. Quart. **14**, 241–255 (1967).

[2] McKean, R. N., "Criteria, Analysis for Military Decisions," RAND Corporation, R–387–PR (Nov. 1964) (edited by E. S. Quade).

[3] Dantzig, George B., *Linear Programming and Extensions* (Princeton University Press, Princeton, N.J., 1963).

ALLOCATION OF CARRIER-BASED ATTACK AIRCRAFT USING NON-LINEAR PROGRAMMING

Edward W. Rice

Naval Explosive Ordnance Disposal Facility

Jerome Bracken

Institute for Defense Analyses

Arthur W. Pennington

Office of the Chief of Naval Operations

ABSTRACT

The paper presents the formulation and several solutions of a model for allocating a fixed number of aircraft to carriers and to missions. The amount of damage that can be inflicted is maximized. A nonseparable concave nonlinear objective function expresses diminishing marginal damage. Linear constraints on aircraft, carrier space, and aircraft availability for missions are included. The model is solved using the sequential unconstrained minimization technique (SUMT). The model is presented in terms of a scenario. Several different exponential damage functions are treated, and S-shaped damage functions are discussed.

I. INTRODUCTION

The efficient allocation of weapons among targets is an interesting and difficult resource allocation problem. Once a threat has been identified, there are many potential ways of dealing with it. In the short run (fixed capital inventory) planners are usually forced to allocate existing weapons systems. Large capital investments and long lead times required for receipt of new systems necessitate such an allocation.

Allocating a fixed weapon inventory can be approached in two ways. The amount of damage that can be inflicted on enemy targets can be maximized subject to a constraint on resources. Alternatively, the cost of inflicting a specified amount of damage can be minimized. The second approach assumes that the specified amount of damage can be attained with available resources.

This paper is concerned with maximizing the amount of damage that can be inflicted with a fixed quantity of resources. Specifically, the allocation of carrier-based attack aircraft to targets is addressed. The allocation is two-fold. Attack aircraft are allocated among carriers, and missions are allocated among targets.

The general approach used in the allocation process proceeds as follows. A model is developed that contains a nonlinear objective function and linear constraints. The nonlinearity represents diminishing marginal damage value. Linearity of the objective function would severely restrict the model, for linearity (constant marginal damage value) would apply for only a small number of missions against a target complex. The first few missions against a target complex may behave linearly, but "overkilling" or overlapping of damage areas is more realistic when extensive damage is considered. Extensive damage value is assumed in developing the objective function of the model.

A scenario is devised. The scenario includes groups of targets and carrier locations relative to these targets. Values for parameters of the model are derived using the scenario. Each target group or set is assigned a military value. This target value is discussed briefly. Given the resource constraints (fixed supply of carriers and aircraft), the model is solved. The solution allocates aircraft to carriers and missions to targets to maximize damage.

The model is exercised in several ways to demonstrate its applicability. Situations where damage progresses nonlinearly with each succeeding mission and situations involving approximations of S-shaped damage functions are solved. Various mixes of target values are used in the exercises.

The models are solved using the sequential unconstrained minimization technique developed by Fiacco and McCormick and described in Reference [3]. References [1] and [2] provide the basic structure for the weapons allocation models developed in the present paper.

II. MODEL

Objective Function

The allocation problem is concerned with maximizing the damage value inflicted on an enemy's target set. The objective function of the nonlinear programming problem expresses the amount of damage done weighted by a target value. The damage that is to be inflicted is constrained by a fixed capital inventory of attack aircraft and carriers. The following indices are defined:

$i = 1, \ldots, m$ is the index of carriers,
$j = 1, \ldots, n$ is the index of attack aircraft, and
$k = 1, \ldots, p$ is the index of targets.

Since the damage value inflicted on the target sets involves certain physical characteristics of the attack aircraft, mission production capability of the aircraft, and the damage each mission causes, the following are defined:

x_{ij} = the number of aircraft of type j assigned to carrier i;

a_{jk} = the fraction of target k (either in numbers of specific targets or target area) left undamaged following a mission by an aircraft of type j; and

y_{ijk} = the number of missions flown against target k, by aircraft of type j, from carrier i.

The fraction of target k left undamaged following missions of aircraft of type j from carrier i is $a_{jk}^{y_{ijk}}$. Since the objective function expresses damage inflicted, it contains the fraction of target damaged, $1 - a_{jk}^{y_{ijk}}$. The damage inflicted by missions of aircraft of all types from carrier i against target k is

$$1 - \prod_{j=1}^{n} a_{jk}^{y_{ijk}}.$$

The damage inflicted by missions of aircraft from all carriers against target k is

$$1 - \prod_{i=1}^{m} \prod_{j=1}^{n} a_{jk}^{y_{ijk}}.$$

The total damage inflicted is the sum of damage to all targets

$$\sum_{k=1}^{p} \left(1 - \prod_{i=1}^{m} \prod_{j=1}^{n} a_{jk}^{y_{ijk}} \right).$$

Note that in this derivation the damage to any individual target set is obtained nonlinearly, but that the total damage is obtained additively or linearly. The nonlinear individual target damage reflects the assumption that there is overlapping of damage done by succeeding missions. The additivity of total damage to all targets indicates the underlying assumption that the target sets are separated such that missions flown against one set will not affect any other set. The final step in developing the objective function is weighting each target set by a value index so the objective function shows the total damage value to the enemy's targets. The objective function to be maximized is

$$\sum_{k=1}^{p} u_k \left(1-\prod_{i=1}^{m}\prod_{j=1}^{n} a_{jk}^{y_{ijk}}\right),$$

where u_k is the military target value of target type k.

The validity of this objective function is increased as the situation tends towards nonselective bombing of large area targets, and is decreased as the situation tends towards bombing a collection of point targets, particularly if they are mobile.

Without the military value term, the units of the objective function are percentage or fraction of the target set either in terms of numbers of specific targets or target area. The representation in numbers or area would depend on the nature of the target set. Close support mission damage would probably be represented by numbers of targets damaged. Damage to a city could be represented by the total area damaged. The military value term can be expressed in a number of ways. If the value term represents the size of a given target set relative to the other target sets, the objective function would be expressed as the fraction of the total target complex damaged. The weighting in this case allocates missions to the larger target sets first. The military value term could also represent the priority assigned to the objectives for a given stage of a conflict. For example, at the beginning of an engagement air superiority may be the principal objective. Those target sets associated with the enemy's air potential would then have a high military value. In this case, the objective function would represent the fraction of the enemy's war potential damaged. The military value term could also be an estimate of the value of the target complex in dollars. In this case the objective function would represent the total damage inflicted in dollar terms.

Constraints

Constraints are associated with the fixed supply of aircraft and carriers. Parameters are defined as follows:

b_j = number of aircraft of type j available;

c_j = carrier space required by each of the type j aircraft;

d_i = space available on carrier i; and

f_{ijk} = the fraction of total missions available in the planning period used in flying one sortie against target k by aircraft of type j from carrier i.

Constraints on the total attack aircraft available are

$$\sum_{i=1}^{m} x_{ij} \leq b_j, \quad j=1, \ldots, n. \tag{1}$$

Carrier space limitations restricting the numbers of all types of aircraft on each type carrier are

$$\sum_{j=1}^{n} c_j x_{ij} \leq d_i, \quad i=1, \ldots, m. \tag{2}$$

Limitations on the number of missions specifying that more missions cannot be flown than can be produced by the available numbers of aircraft are

$$x_{ij} - \sum_{k=1}^{p} f_{ijk} y_{ijk} \geq 0, \qquad i = 1, \ldots, m; \; j = 1, \ldots, n. \tag{3}$$

All variables in the model are subject to nonnegativity restrictions

$$\begin{aligned} & x_{ij} \geq 0, \\ & y_{ijk} \geq 0, \quad i = 1, \ldots, m; j = 1, \ldots, n; k = 1, \ldots, p. \end{aligned} \tag{4}$$

III. DESCRIPTION OF EXAMPLE OF MODEL APPLICATION

Scenario

To demonstrate the application of the model, an arbitrary scenario is devised. The scenario is used to assign military values to targets and to develop parameters required for the model.

Assume that a small country under treaty with the United States is under attack by a neighboring country. The United States decides to intervene to honor the treaty. A carrier task force in the vicinity is ordered into action. A command policy decision is made to honor all political and military restrictions on the limited conflict to be conducted. The immediate operational decision made is to maximize damage value to enemy targets. Decision parameters of the allocation problem are the military values placed on the targets, and decision variables are the allocation of attack aircraft to carriers and missions to targets. The immediate short-run constraint on the mission output is the fixed number of aircraft and carriers. The immediate goals to be achieved are to attain air superiority and to provide close air support to the local troops as requested. The targets established based on the goals and intelligence data are:

1. Airfields
2. Fuel Storage Depots
3. Close Support
4. Shore Interdiction
5. Command and Control

The mission allocation assigners are aware that the air superiority goal for the carrier task force is to be in effect only until the Air Force locates forces in neighboring ground bases. They are also aware that no alternative attack mode is available (no ground artillery or sea-based gunfire). The assigners therefore aspire only to hit enemy airfields and provide close support initially. A few missions are to be directed toward fuel storage depots as an aid to keeping enemy air forces inoperative. Close support targets are valued slightly higher than airfield targets because close support is to be a continuing role. The values listed under Phase 1 in Table 1 are assigned (based on an arbitrary total of 100). An evaluation of results after some time interval indicates that the close support role is being satisfied but that the enemy air force is still flying more missions than anticipated. An analysis indicates that many airfields have been damaged, but the remaining capability is producing many enemy missions. The values are modified to those under Phase 2 in Table 1 to allocate more missions against enemy airfields. During Phases 1 and 2, no values are assigned to target sets 4 and 5 because they are not directly related to immediate goals and because of the limited missions available. Later, the Air Force establishes bases and begins flying missions. The carrier task force role is modified. The close support objective is maintained, a new objective of denying supply of enemy guerrilla troops via remote coastal areas is

added, and the air superiority objective is dropped (transferred to the Air Force). Excess missions are to be scheduled against command and control targets. The new target values assigned appear under Phase 3 of Table 1. These three sets of target values are to be used to demonstrate the application of the model.

TABLE 1. *Military Target Values*

Target set	Phase 1	Phase 2	Phase 3
1. Airfields	40	45	0
2. Fuel storage depots	10	15	0
3. Close support	50	40	45
4. Shore interdiction	0	0	45
5. Command and control	0	0	10

Derivation of Parameters

Figure 1 depicts the scenario devised to permit estimations of parameters of the model. Carrier $i=1$ is assumed to be a conventional carrier of the Midway class (CVA). Carrier $i=2$ is a nuclear carrier (CVAN). Further assume that there are two types of attack aircraft available, the A–6 all-weather attack aircraft ($j=1$), and the smaller A–7 ($j=2$). Using the scenario and aircraft and carrier assumptions, the following parameters are generated:

1. a_{jk}. These values are normally based on the type and weight of ordnance carried, number of targets in the set, and the average number of targets destroyed by each mission. They reflect that only a very small fraction of a target is destroyed by a weapon. The following numbers are used in the model:

$$a_{11}=0.99978 \qquad a_{21}=0.99953$$

$$a_{12}=0.99978 \qquad a_{22}=0.99953$$

$$a_{13}=0.99978 \qquad a_{23}=0.99953$$

$$a_{14}=0.99935 \qquad a_{24}=0.99960$$

$$a_{15}=0.99935 \qquad a_{25}=0.99960$$

2. b_j.

$$b_1=27 \qquad \text{(A–6 Aircraft)}$$

$$b_2=102 \qquad \text{(A–7 Aircraft)}$$

3. c_j.

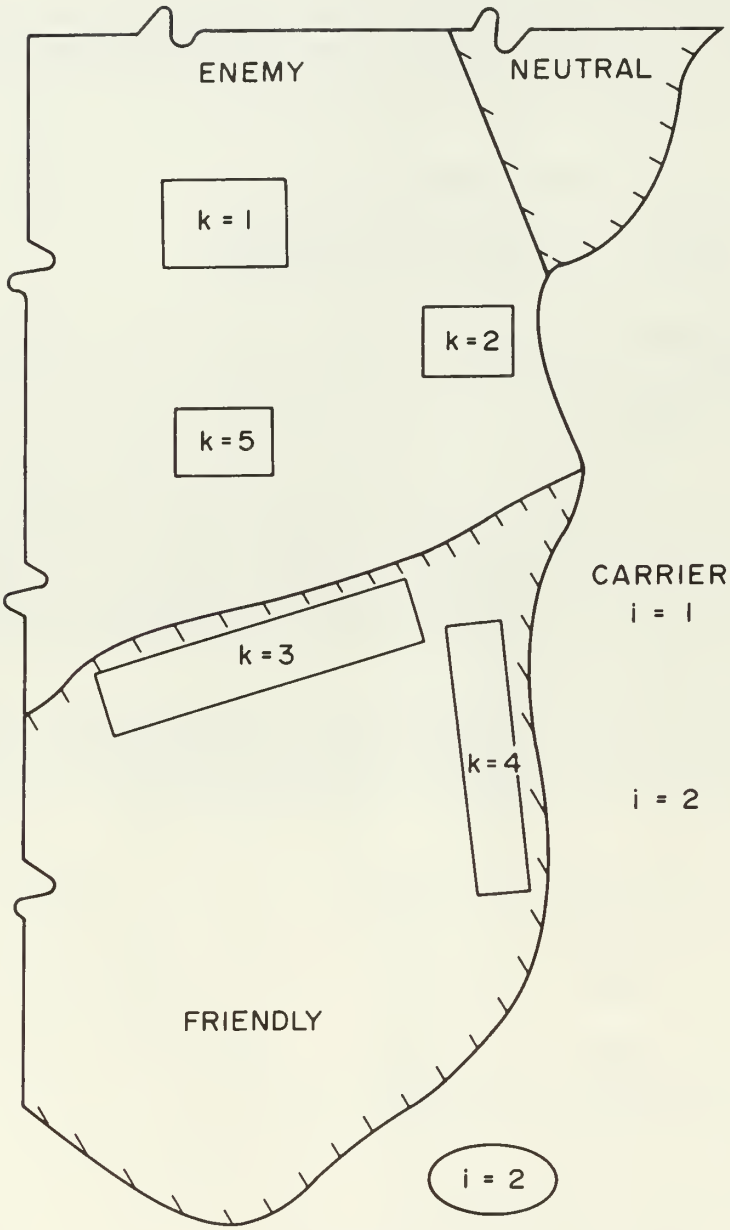
$$c_1=2{,}920 \text{ sq ft} \qquad \text{(A–6 Aircraft)}$$

$$c_2=1{,}770 \text{ sq ft} \qquad \text{(A–7 Aircraft)}$$

4. d_i.

$$d_1=112{,}300 \text{ sq ft} \qquad \text{(CVA)}$$

$$d_2=147{,}100 \text{ sq ft} \qquad \text{(CVAN)}$$



CARRIER	TARGET	DISTANCE
i = 1	k = 1	400
	k = 2	200
	k = 3	400
	k = 4	300
	k = 5	400
i = 2	k = 1	500
	k = 2	400
	k = 3	300
	k = 4	300
	k = 5	400

i = 1

FIGURE 1. Sample problem scenario

5. f_{ijk}. These numbers are determined by the distance to the target and maintenance required between missions. They can be adjusted to reflect damage or attrition effects. The following are used in the model:

$$f_{111} = 0.06452 \qquad f_{211} = 0.06896$$

$$f_{112} = 0.06250 \qquad f_{212} = 0.06452$$

$$f_{113} = 0.06452 \qquad f_{213} = 0.06250$$

$f_{114} = 0.06250$ $f_{214} = 0.06250$

$f_{115} = 0.06452$ $f_{215} = 0.06452$

$f_{121} = 0.05556$ $f_{221} = 0.05882$

$f_{122} = 0.05264$ $f_{222} = 0.05556$

$f_{123} = 0.05556$ $f_{223} = 0.05264$

$f_{124} = 0.05264$ $f_{224} = 0.05264$

$f_{125} = 0.05556$ $f_{225} = 0.05556$

The variables in the model are identified as follows:

Description		*Model variable*
Distribution of aircraft of type j to carrier i		x_{11}
		x_{12}
		x_{21}
		x_{22}
Missions flown from carrier 1 to target k	Using aircraft of type 1	y_{111}
		y_{112}
		y_{113}
		y_{114}
		y_{115}
	Using aircraft of type 2	y_{121}
		y_{122}
		y_{123}
		y_{124}
		y_{125}

Description		*Model variable*
Missions flown from carrier 2 to target k	Using aircraft of type 1	y_{211}
		y_{212}
		y_{213}
		y_{214}
		y_{215}
	Using aircraft of type 2	y_{221}
		y_{222}
		y_{223}
		y_{224}
		y_{225}

Objective Function

Using the parameters just outlined, we identify the components of the objective function are as follows:

Target 1

$$u_1(1-0.99978^{y_{111}}0.99978^{y_{211}}0.99953^{y_{121}}0.99953^{y_{221}})$$

Target 2

$$u_2(1-0.99978^{y_{112}}0.99978^{y_{212}}0.99953^{y_{122}}0.99953^{y_{222}})$$

Target 3

$$u_3(1-0.99978^{y_{113}}0.99978^{y_{213}}0.99953^{y_{123}}0.99953^{y_{223}})$$

Target 4

$$u_4(1-0.99935^{y_{114}}0.99935^{y_{214}}0.99960^{y_{124}}0.99960^{y_{224}})$$

Target 5

$$u_5(1-0.99935^{y_{115}}0.99935^{y_{215}}0.99960^{y_{125}}0.99960^{y_{225}}).$$

Summing these, recalling that the total value sum is 100, and noting that SUMT minimizes the objective function, the following objective function results:

$$\begin{aligned} & u_1\ (0.99978^{y_{111}}\ 0.99978^{y_{211}}\ 0.99953^{y_{121}}\ 0.99953^{y_{221}}) \\ & +u_2\ (0.99978^{y_{112}}\ 0.99978^{y_{212}}\ 0.99953^{y_{122}}\ 0.99953^{y_{222}}) \\ & +u_3\ (0.99978^{y_{113}}\ 0.99978^{y_{213}}\ 0.99953^{y_{123}}\ 0.99953^{y_{223}}) \\ & +u_4\ (0.99935^{y_{114}}\ 0.99935^{y_{214}}\ 0.99960^{y_{124}}\ 0.99960^{y_{224}}) \\ & +u_5\ (0.99935^{y_{115}}\ 0.99935^{y_{215}}\ 0.99960^{y_{125}}\ 0.99960^{y_{225}}). \end{aligned}$$

To facilitate computer programming, the objective function is further modified to include only one a value (0.99978). Missions with other a values are reduced to equivalent missions with an a value of 0.99978 using the following relationship:

$$0.999\mathrm{XX}^{Y}=0.99978^{Z} \text{ or}$$

$$Z=Y\,\frac{\ln\ 0.999\mathrm{XX}}{\ln\ 0.99978}.$$

For example, the $0.99953^{y_{121}+y_{221}}$ damage value against target type 1 is reduced as follows:

$$0.99953^{y_{121}+y_{221}}=0.99978^{Z} \text{ or}$$

$$Z=(y_{121}+y_{221})\,\frac{\ln\ 0.99953}{\ln\ 0.99978}.$$

Having found all mission equivalents, the objective function becomes:

$$u_1\left(0.99978^{y_{111}+y_{211}+[y_{121}+y_{211}]\frac{\ln 0.99953}{\ln 0.99978}}\right)$$

$$+u_2\left(0.99978^{y_{112}+y_{212}+[y_{122}+y_{222}]\frac{\ln 0.99953}{\ln 0.99978}}\right)$$

$$+u_3\left(0.99978^{y_{113}+y_{213}+[y_{123}+y_{223}]\frac{\ln 0.99953}{\ln 0.99978}}\right)$$

$$+u_4\left(0.99978^{[y_{114}+y_{214}]\frac{\ln 0.99935}{\ln 0.99978}+[y_{124}+y_{224}]\frac{\ln 0.99960}{\ln 0.99978}}\right)$$

$$+u_5\left(0.99978^{[y_{115}+y_{215}]\frac{\ln 0.99935}{\ln 0.99978}+[y_{125}+y_{225}]\frac{\ln 0.99960}{\ln 0.99978}}\right)-100.$$

Constraints

The basic constraints are stated as follows:

1. Total aircraft available:

$$27-x_{11}-x_{21}\geq 0$$

$$102-x_{12}-x_{22}\geq 0.$$

2. Carrier space available:

$$112{,}300-2{,}920x_{11}-1{,}770x_{12}\geq 0$$

$$147{,}100-2{,}920x_{21}-1{,}770x_{22}\geq 0.$$

3. Aircraft availability for missions:

(a) $$x_{11}-0.06452y_{111}-0.06250y_{112}-0.06452y_{113}-0.06250y_{114}-0.06452y_{115}\geq 0$$

(b) $$x_{12}-0.05556y_{121}-0.05264y_{122}-0.05556y_{123}-0.05264y_{124}-0.05556y_{125}\geq 0$$

(c) $$x_{21}-0.06896y_{211}-0.06452y_{212}-0.06250y_{213}-0.06250y_{214}-0.06452y_{215}\geq 0$$

(d) $$x_{22}-0.05882y_{221}-0.05556y_{222}-0.05264y_{223}-0.05264y_{224}-0.05556y_{225}\geq 0.$$

IV. SOLUTIONS

Solutions for Exponential Damage Functions

The objective function derived from the sample problem assumes an exponential damage function shown in Figure 2. This type of damage function applies to situations where there is no target hardening. In addition, it assumes that each specific target in the set or each fraction of the target area is of the same importance as any other specific target or fraction of target area. The term importance in this case may mean that the functional ability of a target set is reduced by the same amount if any specific target or fraction of the target area is damaged. The following allocation schemes are solved using the assumption that the exponential damage function in Figure 2 is valid for all target sets.

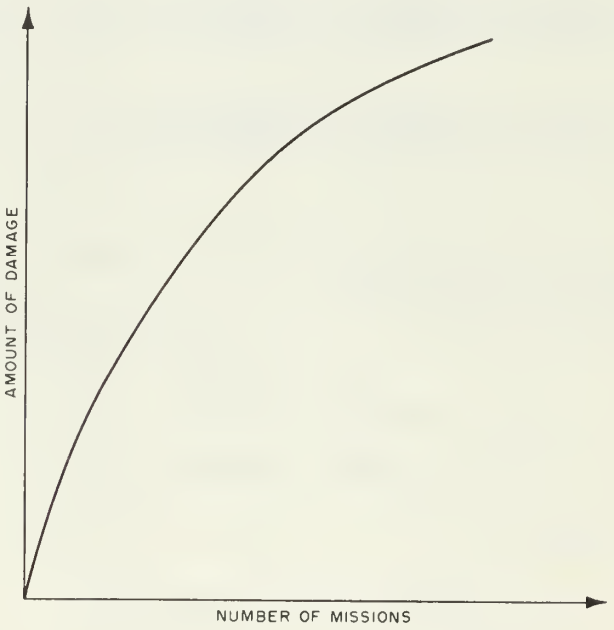


FIGURE 2. Exponential damage function

Allocation Problem 1. The first allocation problem is solved using the Phase 1 target values from Table 1. The SUMT solution is listed under allocation 1 in Table 2. The mission allocation shows that the marginal damage return of initial missions against fuel storage depots is just beginning to be more profitable than additional missions against airfields and close support targets. The value of the objective function is 66.33. Since there is an arbitrary total of 100 target value units, the value of the objective function shows that 33.67 percent of the total target value is damaged. Note that 873 missions are scheduled against airfields and 1,432 missions against close support targets.

Allocation Problem 2. The objective function is modified to include the Phase 2 target values from Table 1. The constraints remain unchanged. The solution is shown under allocation 2 in Table 2. The primary objective of the Phase 2 target values is to determine if small changes in target values can be used to reallocate missions among targets. The results for this allocation scheme indicate that it is possible. The allocation results in decreasing the missions against close support targets and increasing missions against airfields and fuel storage depot targets. This is the reallocation that is desired as

TABLE 2. *SUMT Allocation Solutions*

Missions		Allocation				
		1	2	3	4	5
		Total target value damage (%)				
		33.67	31.27	37.35	31.06	26.92
Aircraft allocation						
	x_{11}	18	4	23	14	8
	x_{12}	33	56	26	40	50
	x_{21}	9	23	4	13	19
	x_{22}	69	46	76	62	52
Mission allocation						
Airfields	y_{111}	250	46	.....	194	107
	y_{121}	584	1,000	.....	701	879
	y_{211}	26	66	.....	30	36
	y_{221}	13	25	.....	14	15
Fuel storage depots	y_{112}	3	3	.....	2	2
	y_{122}	1	2	.....	1	1
	y_{212}	2	4	.....	2	2
	y_{222}	1	2	.....	1	1
Close support	y_{113}	26	14	.....	23	18
	y_{123}	14	10	22	13	11
	y_{213}	106	282	.....	163	252
	y_{223}	1,286	836	1,371	1,164	977
Shore interdiction	y_{114}	.....	.....	358	.....	.....
	y_{124}	.....	.....	472	.....	.....
	y_{214}	.....	.....	68	.....	.....
	y_{224}	.....	.....	66	.....	.....
Command control	y_{115}	.....	.....	.....	.....	.....
	y_{125}	.....	.....	1	.....	.....
	y_{215}	.....	.....	.....	.....	.....
	y_{225}	.....	.....	1	.....	.....
Total missions		2,312	2,290	2,359	2,308	2,301

reflected in the target value changes. Note also that 14 type-1 aircraft are shifted from carrier 1 to carrier 2, and 23 type-2 aircraft are shifted from carrier 2 to carrier 1. In reality an analysis may show that the cost of moving these aircraft and their support functions and equipment would not be worth the increase in damage value return. Perhaps the aircraft allocation would be held constant and the missions reallocated. The optimum aircraft allocation would be determined by using *SUMT* to allocate the aircraft for all possible target value combinations. The aircraft allocation could then be optimized to maximize the overall damage value considering every possible situation. It would be possible to assign a probability weight to each possible value outcome using this scheme.

Allocation Problem 3. The target values listed under Phase 3 of Table 1 are used for Problem 3. The remaining constraints are the same as those used in Problems 1 and 2. The resulting allocation is listed under allocation 3 in Table 2. The solution allocates missions as might be expected except

that more missions are scheduled against close support targets than shore interdiction targets. This is unexpected because the target values for these targets are equal. The explanation lies in the fact that more damage results from missions by aircraft type-2 against close support targets than against shore interdiction targets (see a_{23} and a_{24} values). There are very few missions flown against command and control targets. The marginal return against these targets has just begun to become more profitable than additional missions against other targets. In reality these missions would probably be ignored depending on their location relative to the other targets.

Solution for S-Shaped Damage Function

Consider the case now where the exponential damage function assumption is not valid for all target sets. Normally, some targets are "hardened" or protected. This protection means that the initial missions against the target result in little or no damage to the functional ability of the target. After the protection has been "softened" by some number of missions, the attackers begin to realize an increase in the damage value to the target set with each succeeding mission. The assumption that each specific target or fraction of target area is as important as any other target or fraction of target area is not valid for all target sets. Consider the target sets devised in the sample problem formulation. The exponential function probably does not apply to the airfield target set for several reasons. First of all, the aircraft may be revetted to protect them from air attack. Initial missions would have to concentrate on damaging the revetment before any damage value to the aircraft could be realized. Also, the areas at either end of a runway could be rendered useless without affecting the number of missions generated from the runway. The close support and fuel storage depot target damage may or may not follow an exponential damage function. For the purpose of further allocation problems, it is assumed that airfield target damage does not conform to the exponential damage function of Figure 2, but that fuel storage depot, close support, shore interdiction, and command and control targets do conform.

What sort of function best describes the damage to "hardened" targets or target sets with unequal specific target or area values? To say that no damage value is realized when "softening" a

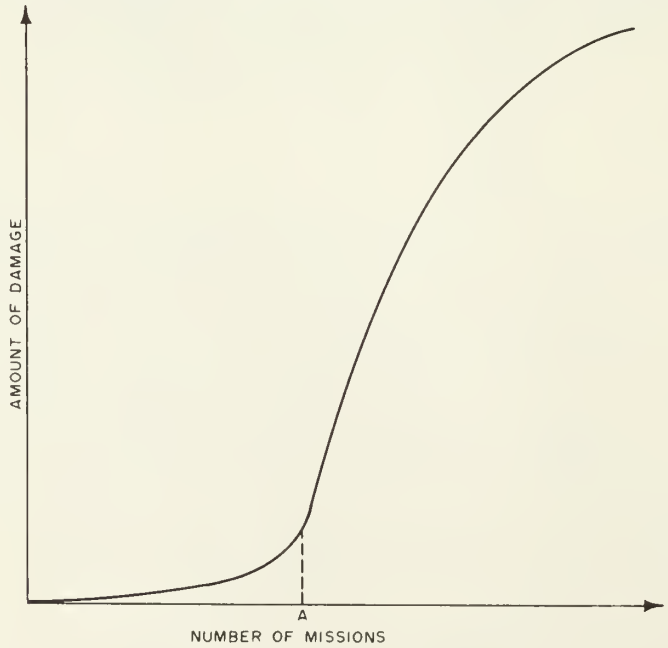


FIGURE 3. S-shaped damage function

target or damaging low valued parts of target sets is not true. Effort must be expended to repair damage to target hardening or target areas. Expense is incurred when replacing low valued parts of target sets. This effort and expense is being diverted from other functions that may have higher values. The S-shaped curve pictured in Figure 3 is usually used to represent damage to target sets that are hardened or contain unequal valued parts. Missions flown up to point A result in little damage. Up to point A the "hardening" is being removed or the low valued targets are being destroyed. Beyond point A, a sharp increase in marginal damage is realized with increasing missions.

The S-shaped damage function presents a problem where the *SUMT* technique is concerned; the convexity necessary to insure that a local optimum is the global optimum being sought is violated. By using several starting points, convergence to the same solution is sometimes used to indicate that a global solution may have been achieved. If starting points are located both to the left and right of A in Figure 3, perhaps this criterion is adequate. It is difficult to determine, however, if a feasible starting point is in the area in question. *SUMT*, then, can possibly be used to solve certain nonconvex nonlinear programming problems providing a history of common solutions converging from various starting points can be compiled. The convexity criterion is not violated in this paper. Approximating techniques of S-shaped damage functions that do not result in nonconvex regions are examined.

Alternate Curve Form

One way of approximating S-shaped damage functions without violating the convexity criterion is to approximate only the part following the knee (right of A in Figure 3) and permit an intercept with the mission axis other than at the origin. Figure 4 illustrates the approximation graphically.

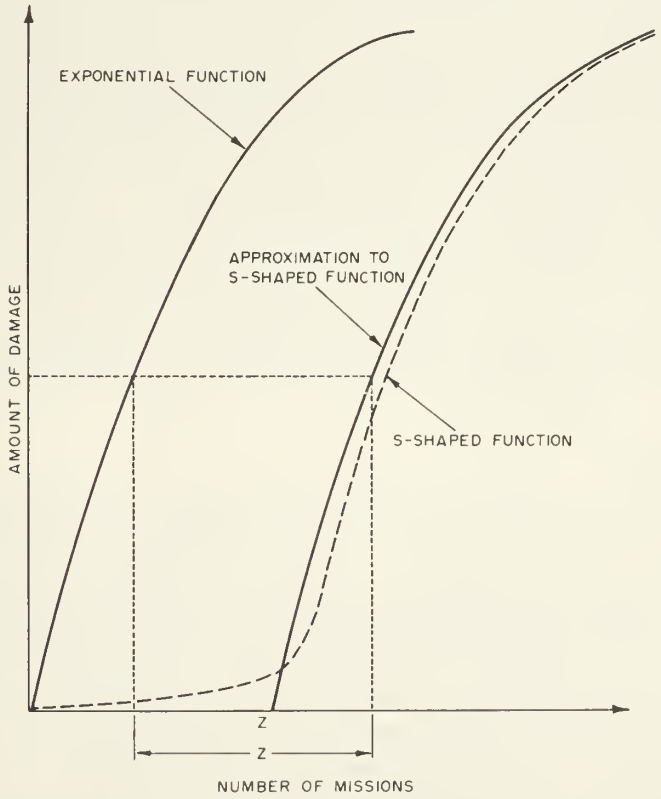


FIGURE 4. Approximation of S-shaped damage function and an exponential damage function

The curve form used to approximate the S-shaped curve following the knee is the exponential curve shown in Figure 3 displaced from the origin by some number of missions. The knee of the S-shaped function is assumed to occur at Z missions (this number is to be flown before any significant damage results). The exponential curve and the curve approximating the S-shaped curve are shown. Note that to obtain any given amount of damage Z more missions have to be flown on the curve approximating the S-shaped curve than on the exponential curve. The damage to airfields using the S-curve approximation is modified to:

$$u_1\left(.99978^{\,y_{111}+y_{211}+[y_{121}+y_{221}]\frac{\ln .99953}{\ln .99978}-Z}\right).$$

Allocation Problem 4. The objective function of Problem 1 is modified to include the S-shaped curve approximation. The remainder of Problem 1 remains unchanged. The allocation solution is listed under allocation 4 in Table 2. Because 873 missions are scheduled against airfields in Problem 1 and the knee of the S-shaped curve is assumed to occur at 400 missions in this problem, the mission allocation is not expected to change a great deal. An increase of 76 missions against airfields results. The slope of the damage value curve is greater at any number of missions over 400 for the Problem 4 damage function than for the Problem 1 damage function. The increased slope gives a higher marginal return with increased missions which explains the mission increase. The value of the objective function is lower for any given number of missions in Problem 4 than in Problem 1. Note that the Problem 4 solution calls for a different aircraft distribution than Problem 1.

Allocation Problem 5. Problem 4 is repeated for an S-shaped curve approximation with the knee occurring at 1,000 missions. The solution is listed under allocation 5 in Table 2. The aircraft distribution is again altered, and 1,037 missions are scheduled against airfields. The damage to target value is further reduced. This solution points to a potential weakness in the S-shaped curve approximation scheme. The fact that the knee of the S-shaped curve occurs at 1,000 missions indicates that at least 1,000 missions are needed to remove target hardening. Only 37 missions are scheduled beyond removal of the hardening because of the value weights. About 19 percent of the airfield value units are damaged. In Problem 4, 949 missions result in 25 percent damage and in Problem 1, 873 missions result in 29 percent damage. As the missions at the knee of the S-shaped curve increase, the damage value decreases although the missions scheduled increase. At some point it is profitable to schedule missions against other targets and none against airfields. The S-shaped curve approximation technique does not allow exercising this option.

Integer Requirements

Problems 4 and 5 do not allow the option of not scheduling certain missions when it is profitable not to schedule them. The S-shaped curve approximation technique is valuable for certain conditions, but it does not cover all possible conditions. A reexamination of Figure 4 points out that it is not profitable to schedule any number of missions less than A unless none are scheduled at all. It is desirable to schedule some number in excess of A that will result in some damage to the target beyond breaking down the "hardening." It is desirable to schedule *either* no missions *or* some number greater than A. This statement means that the S-shaped curve can be approximated by the curves used in Problems 4 and 5, thereby keeping the convexity property of the feasible region. However, by including the origin in the feasible region, a discontinuity is introduced. The functions are no longer twice continuously differentiable. The "either/or" choice posed points to a mixed integer programming model.

The fact that not all damage functions pose the "either/or" choice further indicates a mixed integer-continuous variable model. The literature includes solution techniques for integer linear programming problems, but limited computational experience exists. Nonlinear integer programming solutions are quite limited. There currently is no proven general nonlinear integer programming solution technique. If one existed, it would be useful for this problem.

REFERENCES

[1] Bracken, J. and G. P. McCormick, *Selected Applications of Nonlinear Programming* (Wiley, N.Y., 1968).

[2] Bracken, J. and A. W. Pennington, "Nonlinear Programming Model for Maximizing Target Damage Value by Optimal Allocation of Bomber Aircraft to Carriers and Sorties to Targets," Navy Force Structure Study Working Paper WP–13, Institute for Defense Analyses, March 1968.

[3] Fiacco, A. V. and G. P. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques* (Wiley, N.Y., 1968).

DYNAMIC PROGRAMMING APPROACH TO THE OPTIMIZATION OF NAVAL AIRCRAFT REWORK AND REPLACEMENT POLICIES*

Arnold Neil Schwartz

Center for Naval Analyses

LCDR James A. Sheler

Center for Naval Analyses
Officer Study Group

CDR Carl R. Cooper

Naval Air Systems Command

ABSTRACT

This paper describes a method for determining optimal repair and replacement policies for aircraft, with specific reference to the F-4. The objective of the analysis is to choose the set of policies from all possible alternatives over a finite planning horizon which minimizes the cost of operations. A dynamic program is presented which seeks an optimal path through a series of decision periods, when each period begins with the choice of keeping an aircraft, reworking it before further operation, or buying a new one. We do not consider changes in technology. Therefore, when a replacement does occur, it is made with a similar aircraft. Multivariate statistical techniques are used to estimate the relevant costs as a function of age, and time since last rework.

I. INTRODUCTION

There is a strong effort within the Navy today to improve the aircraft maintenance program. This interest resulted from the rapid rise in the number of naval aircraft unavailable for squadron use during the current Vietnam conflict for reasons of maintenance. In 1966 the Chief of Naval Operations requested the Center for Naval Analyses to investigate the Naval maintenance program for the purpose of reducing the number of aircraft out of service. Initial findings were so appealing that the scope of the study was expanded. One ensuing task was for the development of a rationale for retirement of aircraft based on economic considerations.

Previously, for the most part, aircraft rework and replacement decisions involved historical evidence, planning factor costs, and immediate operational requirements. Selection of these elements of decisions were to some extent arbitrary, and the effects of time on these elements were rarely considered. Nevertheless, the desirability for a methodology to predict the cost impact of present and future decisions had been recognized, and the results of CNA's preliminary study attested to the feasibility of developing such methodology.

---

*This paper is a revision of INS Study 20, "A Dynamic Programming Approach to the Optimization of Naval Aircraft Rework and Replacement Policies" conducted under contract No. 0014-14-68-A-0091 at the Institute of Naval Studies, Center for Naval Analyses of the University of Rochester. No part of this paper necessarily represents the opinion of the Department of the Navy.

Principal Factors

There are three principal factors which interact to establish the service life of an aircraft:

- Material condition.
- Mission effectiveness.
- Economic considerations.

Material condition. The material condition of an aircraft tends to decline as the aircraft ages: major components of the aircraft approach a fatigue life limit, unscheduled maintenance actions occur more frequently, and the underlying mechanical wear and tear factors along with corrosion effects all accumulate to produce the aging effect. This effect can be deterred somewhat through an intensive maintenance program. Nevertheless, it is rarely appropriate to combat the aging effect "at all costs," so in reality some aging does take place even if an economically optimal maintenance program is operative. The aging effect manifests itself, as the aircraft ages, as a decreasing readiness rate, which influences mission effectiveness adversely and thereby affects the aircraft service life.

Mission effectiveness. In general, owing to its material condition, the ability of an aircraft to perform a stipulated mission declines with the passing of time. Moreover, the succession of stipulated mission requirements tends to be increasingly difficult for a type-model-series of aircraft to achieve. In sum, the mission effectiveness of an aircraft declines as the gap widens between mission requirements and attainable aircraft performance. Consequently, this mission effectiveness factor becomes increasingly important throughout time in its effect on aircraft service life determination. In fact, under some circumstances it could dominate the other two factors—material condition and economic cost—but ordinarily all three factors influence that determination.

Economic considerations. Many variables contribute to a given level of aircraft mission effectiveness. Some of the significant ones are: (1) depth of maintenance, (2) length of the planned maintenance cycle, (3) aircraft type, and (4) aircraft age. Resources available to attain any level of aircraft mission effectiveness are limited. Axiomatically the controllable factors should be adjusted to minimize resource use compatibly with attainment of desired effectiveness for a given planning horizon. In that way economics influences the aircraft maintenance plan and the service life determination. In the past, most carrier-based aircraft were assigned a seven-year service life. This decision was based primarily on subjective judgment, because an effective planning tool was lacking for determining the service life of an aircraft based on a proper integrated consideration of material condition, mission effectiveness, and economics.

Maintenance Concepts

The Navy's concept of aircraft maintenance has long included a periodic processing at industrial facilities which have capabilities and skills exceeding those of the fleet maintenance activities. This periodic high-level maintenance became known as depot level maintenance. From the 1940's through the 1950's and 1960's there was an evolutionary development of this depot level maintenance from an overhaul program, to an Interim Rework program, and then to a PAR (Progressive Aircraft Rework) program.

Aircraft overhaul. Overhaul consists of a complete disassembly of an aircraft to permit inspection of all operating components and all basic elements of the aircraft structure. This is followed by repair, replacement or servicing, the incorporation of changes required by technical directives, and flight test to a mission ready status, i.e., all systems in the aircraft in normal operating condition. An integral

part of the overhaul concept involves a change of aircraft custody from the fleet operating unit to the shore establishment during overhaul. Simultaneously, a replacement aircraft which has just completed overhaul is delivered from the shore establishment to the fleet unit. Consequently, the recipient becomes accustomed to receiving a virtually remanufactured aircraft. This overhaul concept is still operative for a limited number of naval aircraft.

The length of service tour between overhauls varied by aircraft model. During the 1940's, however, when overhaul applied to virtually all naval aircraft, the service tour averaged 26 months. This time interval was the prime reason for the next evolution in the depot maintenance concept. About 1952, with the introduction in the fleet of the new jet aircraft, aircraft technology was advancing so rapidly that aircraft needed frequent updating. In the 26-month period from one overhaul to another, a wide gap developed between the actual configuration/capabilities of the aircraft and the potential level that technology would permit. Such a gap was not acceptable to the fleet, and the Interim Rework concept evolved. With Interim Rework in operation, the overhaul and the 26-month service tour were both maintained, but in the middle of the tour a 30-day interim rework was performed at the industrial facility. The interim rework concentrated on updating the aircraft through modifications.

Aircraft interim rework. The Interim Rework concept met the fleet need for frequent updating of the aircraft configuration, but generated an adverse effect because of increased out-of-service time. The *P*rogressive *A*ircraft *R*ework (PAR) program was initiated to obviate this new problem.

Progressive aircraft rework. The PAR concept became possible owing to rapid technological advances in developing basic aircraft materials. The result was improvements in basic structure which precluded the need to rework an airplane to the exhaustive depth of earlier years. For example, the useful life of aircraft wiring was so greatly increased that the old practice of rewiring during overhaul was no longer required. The technological replacement of "wood, dope, and glue" by metal, sophisticated fasteners, and honeycomb reduced the scope of active processing and turned the PAR emphasis away from routine remanufacturing tasks toward a detailed inspection of the aircraft and correction of defects discovered during inspection.

PAR is based on a precept of adjusting rework content and frequency as necessary to preclude the need for overhaul, and to assure, within high confidence limits, continuance of a material condition which will sustain the aircraft through a subsequent operating tour. To minimize out-of-service time, fleet maintenance actions are avoided at the depot. If minor aircraft discrepancies are discovered during rework, they are left uncorrected if they do not interfere with the rework process or affect the safe flight of the aircraft. This policy on minor maintenance items is flexible. The extent to which this policy is implemented reflects a compromise between minimal in-process time and the natural fleet desire to accept an aircraft with all systems in a normal operating condition.

In tracing out the evolution of Navy depot-level maintenance concepts, several factors influencing the maintenance decisions are identified, but economic costs associated with each maintenance concept are conspicuous by their absence. This shortcoming in decision making under maintenance policies, existing and abandoned, is dealt with by providing a dynamic program which introduces a systematic consideration of costs for all alternative decisions that might be undertaken.

II. THE PROBLEM

The problems to be solved are:

1. The age at which the F–4A should be replaced with the F–4J.
2. The determination of an optimal rework schedule.

III. METHOD OF SOLUTION

Suppose that at the beginning of a planning period the decision-maker can choose to keep an aircraft operating or purchase a new one. If we consider N_o planning periods, then the total number of alternatives available over this planning horizon is 2^{N_o}. In general, the total number of alternatives is equal to the number of choices available at the beginning of a planning period raised to the power N_o, the number of planning periods. If there are two alternatives, the decision to keep denoted by (K), and the decision to purchase denoted by (P), and the planning horizon is three periods, then the total number of alternatives will be eight. The alternatives are:

1. (P_1, P_2, P_3)
2. (K_1, K_2, K_3)
3. (P_1, K_2, P_3)
4. (P_1, P_2, K_3)
5. (K_1, P_2, P_3)
6. (K_1, P_2, K_3)
7. (P_1, K_2, K_3)
8. (K_1, K_2, P_3)

The costs associated with a decision to keep an aircraft operating for a planning period are the maintenance costs, consisting of labor and material costs, and an imputed cost of downtime. If the decision is to purchase, then the costs will include the purchase price of the replacement aircraft less its residual value plus the maintenance costs of a new aircraft for the planning period. It is assumed that all costs are incurred at the beginning of a planning period and that these costs can be related to the age of the aircraft.

If C_i^J represents the cost in the ith period for the Jth alternative and N_0 is the number of periods, then the decision rule is to select that alternative, J^M, which results in minimum costs:

$$\text{Cost minimum} = C_1^{J^M} + \sum_{i=2}^{N_o} \frac{C_i^{J^M}}{(1+r)^{(i-1)}} \tag{1}$$

where r is the appropriate discount rate. The least cost alternative will indicate the period(s) when purchases should be made.

It is evident that as both the number of choices available at the beginning of a planning period and the number of periods increase, the number of alternatives that have to be evaluated become extremely large. The dynamic programming algorithm derived by Bellman and Dreyfus [2], illustrated below, is an efficient method for selecting the least cost alternative with the use of a digital computer. It was used to solve the two problems considered here. In the appendix we show the equivalence of the "present value" solution, using Equation (1), with the dynamic programming solution.

The example with two choices available at the beginning of a planning period was used to simplify the discussion of the decision rule that is used to determine when to purchase a new aircraft. For purposes of determining optimal replacement and rework policies, a third choice is added. The decision-maker can now choose to keep (K), purchase (P), or rework and continue operating (R). If the decision is to rework, then the costs will include the cost of a Progressive Aircraft Rework plus the maintenance cost of the reworked aircraft for the portion of the planning period that the aircraft is not in rework.

The dynamic programming formulation consists of the following set of recurrence relations:

$$(2)\qquad f_N(t_1, t_2) = \text{Minimum} \begin{cases} P\colon U_N(o, o) + C_N(t_1) + \dfrac{1}{(1+r)} f_{N+1}(1, 1) \\ K\colon U_N(t_1, t_2) + \dfrac{1}{(1+r)} f_{N+1}(t_1+1, t_2+1) \\ R\colon U_N(t_1, o) + R_N(t_1, t_2) + \dfrac{1}{(1+r)} f_{N+1}(t_1+1, 1), \end{cases}$$

where

. t_1 is the age of the aircraft,

t_2 is tour length,

r is the discount rate,

$f_N(t_1, t_2)$ is the cost at year N of the overall cost of an aircraft where an optimal replacement policy is employed for the remainder of the process,

$C_N(t_1)$ is the net replacement cost as a function of age,

$R_N(t_1, t_2)$ is the cost of a Progressive Aircraft Rework as a function of age and tour length,

$U_N(t_1, t_2)$ is the maintenance costs as a function of age and tour length, and

N_o is total number of periods being considered. Because the process lasts N_o stages and then stops, $f_{N_o+1}(t_1, t_2) \equiv 0$.

This algorithm is begun by evaluating all admissible values of the function $f_{N_o}(t_1, t_2)$ in the last period, and then using these results to determine all admissible values of the function $f_{N_o-1}(t_1, t_2)$. This procedure continues through to the first period where the minimum cost of the optimal policy is determined. Once the algorithm is completed, the optimal path can be traced out by following the minimum cost decisions beginning with the first period. This indicates the age at which the aircraft should be replaced and the age at which the aircraft should be reworked. In the appendix we show how this algorithm is used in the solution of a simple problem where the decision-maker can choose to keep or replace over a planning horizon of three periods.

A computer program capable of evaluating 80 planning periods was written for this algorithm by Mr. William Pierce of the Center for Naval Analyses. The program is described in [8]. The program has been programmed in 3400 FORTRAN for use on a Control Data Corporation Model 3400 computer, and requires approximately 20,000 words of storage. The running time for the program is about 28 minutes.

Flexibility was added to the program by allowing the decision-maker to suppress alternatives that are not available to him along the optimal path. For example, if the decision at the beginning of the Nth period is to purchase, but no funds will be available at that time for procurement, it is possible to suppress the purchase decision at this period and continue on a sub-optimal path. The cost of deviating from the optimal path can readily be determined. If engineering considerations require certain types of aircraft to be reworked within specified intervals, the decision-maker can force reworks even though the optimal path indicates that reworks do not occur. Again, the cost of deviating from the optimal path will be determined.

IV. ASSUMPTIONS AND LIMITATIONS

Assumptions

1. A planning period will be a calendar quarter, and the number of planning periods will be 80.
2. The annual discount will be 10 percent.
3. When estimates are derived with the use of regression equations, the relationships will be valid beyond the observed range of data.
4. The functions used for estimating costs are the same for the F–4A and the F–4J.

Limitations

It is assumed that the process stops at the end of the 80th period; this implies that all costs beyond the 80th period are zero. It is possible that 80 periods are not sufficient to indicate a recycling of policies; we are not in a position at this time to fully evaluate this end effect.

V. DEVELOPMENT OF COST ESTIMATES

The Regression Model Used in the Cost Analyses

The regression model we used to estimate the cost of a Progressive Aircraft Rework, $R_N(t_1, t_2)$, and the cost of unscheduled maintenance, $U_N(t_1, t_2)$, combined arithmetic and logarithmic variables. The function has the following form:

$$\text{Log}_e Y_i = \alpha + \sum_{j=1}^{n} B_j \text{Log}_e X_{ji} + \sum_{j=1}^{m} \alpha_j W_{ji} + \text{Log}_e \epsilon_i, \qquad i = 1, \ldots, N. \tag{3}$$

This relationship can also be conveniently written as a multiplicative function:

$$Y_i = e^{\alpha} \left[\prod_{j=1}^{n} X_{ji}^{B_j} \right] \left[e^{\sum_{j=1}^{m} \alpha_j W_{ji}} \right] \epsilon_i, \qquad i = 1, \ldots, N \tag{4}$$

where

N is the number of observations

n is the number of logarithmic variables

$B_j = \dfrac{\partial Y}{\partial X_j}\left(\dfrac{X_j}{Y}\right)$ is the percentage change in Y for a given percentage change in X_j, $j = 1, \ldots, n$,

m is the number of arithmetic variables. If the arithmetic variable is a dummy variable with discrete values 0, and 1, then

$e^{\alpha_j W_j}$ is the constant percentage multiplier e^{a_j} when $W_j = 1$, and

ϵ_i is the disturbance.

Estimating the Cost of a Progressive Aircraft Rework as a Function of Age and Tour Length for the F–4

Estimating the man-hours expended for a progressive aircraft rework. It is reasonable to expect the following relationships to exist between man-hours and various measures of flight activity and age:

- Aircraft with more flight hours and arrested landings during a tour require more maintenance man-hours.
- Aircraft with longer tour lengths will require more maintenance man-hours.

- Older aircraft require more maintenance man-hours.
- Maintenance man-hours differ by assignment, i.e., aircraft operating in a combat environment require more maintenance than similar aircraft used in training.

Statistical analysis. Data was collected from individual aircraft undergoing standard reworks at the Naval Air Rework Facility at North Island during the period October 1963 to March 1967. A log-linear regression equation was used to relate man-hours expended on a Progressive Aircraft Rework to:

- Number of arrested landings.
- Age.
- Tour length.
- Flight hours.
- Custodian.*

The regression equation is summarized as follows:

(5)
$$\begin{aligned}\text{Log}_e\,(\text{man-hours}) &= 7.8401 \\ &+ \underset{(3.4)}{0.1645}\ \text{Log}_e\,(\text{tour length}) \\ &+ \underset{(12.8)}{0.2523}\ \text{Log}_e\,(\text{age}) \\ &+ \underset{(6.4)}{0.0006}\ (\text{arrested landings}) \\ &- \underset{(6.2)}{0.0646}\ \text{Log}_e\,(\text{flight hours}) \\ &+ \underset{(0.8)}{0.0184}\ (\text{if training}) \\ &- \underset{(1.4)}{0.0369}\ (\text{if Marine}) \\ \text{Correlation coefficient} &= 0.83 \\ \text{Standard error of estimate} &= 0.1182 \\ \text{Degrees of freedom} &= 210\end{aligned}$$

The t values for the coefficients, which are calculated by dividing the coefficient by its standard error, are shown below the coefficients. An arithmetic variable operates as a multiplier of the form $e^{a_j W_j}$, while the coefficient of the logarithmic variable shows the proportional change in the dependent variable for a proportional change in the independent variable. For example, Equation (5) indicates that a 10 percent change in tour length will increase man-hours by 1.645 percent. Marine aircraft require approximately 4 percent fewer man-hours than Navy-deployed aircraft.

An imputed cost of downtime equal to the estimated cost of an aircraft day times the expected days-in-process is added to the cost of a Progressive Aircraft Rework. The days-in-process is estimated with the following equation:

(6)
$$\begin{aligned}\text{Log}_e\,(\text{days in process}) &= -1.9288 \\ &+ \underset{(8.3)}{0.6526}\ \text{Log}_e\,(\text{man-hours}) \\ \text{Correlation coefficient} &= 0.45 \\ \text{Standard error of estimate} &= 0.2538 \\ \text{Degrees of freedom} &= 210\end{aligned}$$

*Dummy variables were used to distinguish Navy-deployed aircraft from Marine and training aircraft.

Estimating material costs for a progressive aircraft rework. The accounting procedure used by the Naval Air Rework Facilities separates material costs into two categories: Navy Industrial Fund Material, and Government Furnished Material. These costs were collected by individual aircraft that underwent standard reworks at the Naval Air Rework Facility located at North Island. The data is summarized in Table I.

TABLE I. *Means and Standard Deviations of Navy Industrial Fund Material and Government Furnished Material For F–4 Progressive Aircraft Reworks*

Variable	Means and standard deviations
Navy industrial fund material.....	$3565. (1117)
Government furnished material.....	$23,718 (12,962)
Sample size.....	147
Time period.....	Oct. 1963–Jan. 1967

Statistical analysis. Using a log-linear equation to relate Government Furnished Material to the same variables that were specified for Equation (3), we have:

$$
\begin{aligned}
(7)\qquad \text{Log}_e\,(\text{government furnished material}) &= 7.98 \\
&\quad +0.7451\ \text{Log}_e\,(\text{tour length}) \\
&\qquad (1.85) \\
&\quad +0.1461\ \text{Log}_e\,(\text{age}) \\
&\qquad (0.82) \\
&\quad +0.0008\ (\text{arrested landing}) \\
&\qquad (1.46) \\
&\quad -0.3447\ \text{Log}_e\,(\text{flight hours}) \\
&\qquad (2.26) \\
&\quad -0.3974\ (\text{if Marine}) \\
&\qquad (2.00) \\
&\quad -0.1815\ (\text{if deployed}) \\
&\qquad (1.18) \\
&\quad +0.0337\ (\text{time}) \\
&\qquad (4.5) \\
\text{Correlation coefficient} &= 0.54 \\
\text{Standard error of estimate} &= 0.3015 \\
\text{Degrees of freedom} &= 139.
\end{aligned}
$$

A trend variable, time, was included because material costs are measured in dollars. This could be interpreted as a proxy for price and specification changes that have occurred during the observed period.

Navy Industrial Fund material costs were found to be relatively stable over the period observed. Therefore, no significant relationship was found between these costs and our various measures of flight activity and age. When estimating the cost of a Progressive Aircraft Rework, the average cost of Navy Industrial Fund material will be used.

We can now proceed to estimate the cost of a Progressive Aircraft Rework with the use of Equations (5), (6), and (7). Since we are interested in the cost in relation to age and tour length, we can

collapse our multidimensional relationship into three dimensions by holding all other variables constant at the mean. Equations (5), (6), (7) now have the following form:

$$\text{Man-hours} = e^{7.6043}(\text{age})^{0.2523}(\text{tour length})^{0.1645} \tag{8}$$

$$\text{Government furnished material} = e^{8.0191}(\text{age})^{0.1461}(\text{tour length})^{0.7451} \tag{9}$$

$$\text{Days in process} = e^{-1.9288}(\text{man-hours})^{0.6526}. \tag{10}$$

Using \$4.60 per hour as a direct labor charge and \$2.10 per hour as an indirect charge for applied production expense, we can estimate the man-hour cost by multiplying Equation (8) by \$6.70.* Added to this man-hour cost is the Government Furnished Material from Equation (9), Navy Industrial Fund Material of \$3,560, \$17,000 for General Expense, and an imputed cost of downtime. The imputed cost of downtime is derived by multiplying the number of days in process by \$5,940, which is the cost of an aircraft day.

The final equation for the cost of a Progressive Aircraft Rework, given age, t_1, and tour length, t_2, will be:

$$R_N(t_1, t_2) = (e^{7.6043} t_1^{0.2523} t_2^{0.1645})\ \$6.70 + e^{8.0191} t_1^{0.1461} t_2^{0.7451} + (e^{-1.9288}\ (\text{man-hours}\ (t_1, t_2))^{0.6526})\ \$5{,}940 + \$17{,}000 + \$3{,}560. \tag{11}$$

Figure 1 shows the estimated total cost of a Progressive Aircraft Rework as a function of age and tour length. Figure 2 shows the days in process as a function of man-hours.

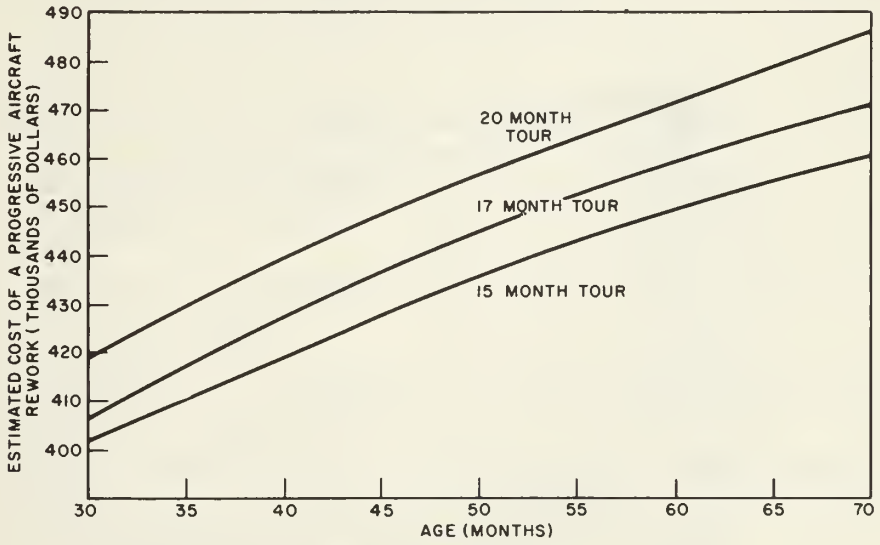


FIGURE 1. The estimated cost of a progressive aircraft rework as a function of age and tour length (including imputed cost of downtime).

*This rate was derived from North Island production and financial statements for the fourth quarter FY–1966.

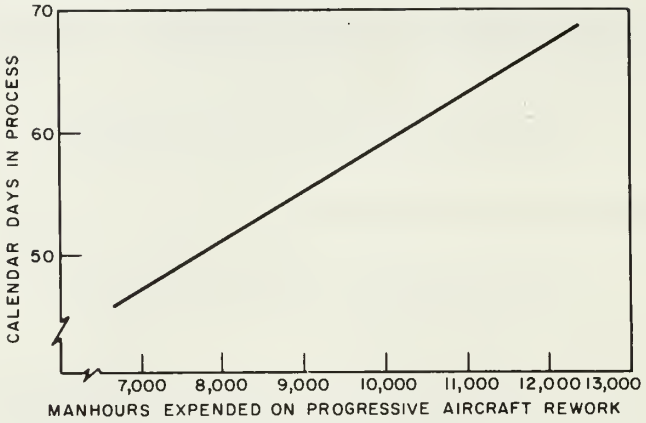


FIGURE 2. Estimated calendar days in process as a function of manhours expended on progressive aircraft rework.

Estimating Replacement Cost as a Function of Aircraft Age for the F-4

The general form of the equation that is used to estimate replacement cost is:

$$\text{Replacement cost} = C(1 - ke^{-\alpha' t}),$$

where

C = Purchase price of the new plane,

T = Age,

α' = Value of the coefficient for a specified percentage rate of annual decline in the salvage value, and

k = Fraction of C remaining as trade-in value after purchasing.

Figure 3* shows the salvage or residual value as a function of age for specified rates of annual decline; Table II shows various values of the coefficient, α', for an arbitrary range of rates of annual decline.

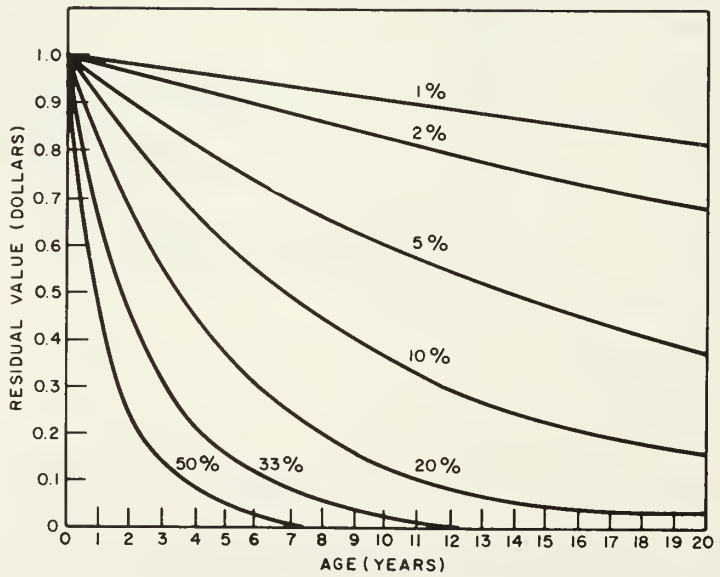


FIGURE 3. Residual values related to various annual rates of decline.

*Extracted from [1].

TABLE II*. *Annual Percentage Rates of Change and Corresponding Values For Coefficients Expressing That Rate of Change*

Percentage rate of annual decline	α'
0.00	0.00
0.01	0.00
0.02	0.02
0.03	0.03
0.04	0.04
0.05	0.05
0.06	0.06
0.07	0.07
0.08	0.08
0.09	0.09
0.10	0.10
0.11	0.12
0.12	0.13
0.14	0.15
0.20	0.22
0.25	0.30
0.30	0.35
0.33	0.40
0.40	0.50
0.45	0.60
0.50	0.70
0.65	1.00
0.70	1.20
0.80	1.60
0.90	2.30
0.95	3.00
0.999	6.00

Estimating Maintenance Costs as a Function of Age and Time Since Last Rework For The F–4

The maintenance performed at the squadron for aircraft consists of the following:

- Unscheduled maintenance—labor and material expended repairing random failures in the aircraft.
- Support maintenance—labor expended in nonrepair actions, such as washing the aircraft, preparing it for flight, corrosion control, etc.
- Scheduled maintenance—labor and material expended as necessary inspecting and repairing specified aircraft components on a periodic basis; scheduled maintenance, which is also called a calendar inspection, occurs every 30 weeks for the F–4.
- Field modification—labor and material expended making engineering changes by field teams sent from the appropriate rework facilities; field teams are used when the work exceeds the capability of the squadron.

Estimating unscheduled maintenance man-hours. Unscheduled direct maintenance man-hours were collected from individual aircraft on a quarterly basis from VF–121. Assuming that older air-

*Extracted from [1].

craft require more maintenance man-hours at the squadron and that maintenance man-hours will increase with the time since last rework, we regressed direct maintenance man-hours on age and time since last rework. This produced the following relationship:

$$\text{(12)}\quad \text{Log}_e(\text{unscheduled maintenance man-hours}) = \underset{(2.4)}{6.5515} + 0.1401\ \text{log}_e(\text{time since last rework}) + \underset{(1.0)}{0.0443}\ \text{log}_e(\text{age})$$

$$\text{Correlation coefficient} = 0.52$$
$$\text{Standard error of estimate} = 0.2165$$
$$\text{Degrees of freedom} = 21.$$

Estimating support man-hours. Support man-hours are reported only for the squadron. The squadron allocates the support man-hours to individual aircraft on the basis of its flight activity. No attempt was made to relate support man-hours to age and time since last rework because the man-hours reported by individual aircraft were not the actual hours expended. Therefore, for all future cost calculations only the average support man-hours per F–4 will be used; the average support man-hours per F–4 for the month of June 1967 for VF–121 was 503.

Estimating man-hours for calendar inspections. Direct maintenance man-hours expended on 19 F–4's for calendar inspections during the period October 1966 to June 1967 were collected from VF–121. It was not possible to relate these man-hours to age and time since last rework because of the insufficient number of observations obtained from this sample of 19. The average man-hours per calendar inspection is 805, and the number of days the aircraft is down is approximately 8.

Estimating man-hours for field team modifications. The man-hours expended on individual F–4 aircraft for field modification were collected from the Naval Air Rework Facility located at North Island for the period October 1965 to February 1968. Since man-hours expended for field team modifications are a function of the time since last rework, regression analysis was used to estimate the functional form of the equation. The equation is:

$$\text{(13)}\quad \text{Log}_e(\text{man-hours}) = 4.0075 + \underset{(3.84)}{0.2699}\ \text{Log}_e(\text{time since last rework})$$

$$\text{Correlation Coefficient} = 0.20$$
$$\text{Standard error of Estimate} = 1.2048$$
$$\text{Degrees of freedom} = 355.$$

An imputed cost of downtime will be added to the maintenance costs. Therefore, the next section discusses the procedure used for estimating the time the aircraft is not operationally ready as a function of age and time since last rework.

Estimating the percentage of time the aircraft is not operationally ready. The percentage of time the aircraft is not available for use is composed of *N*ot *O*perationally *R*eady *M*aintenance (NORM) and *N*ot *O*perationally *R*eady *S*upply (NORS). These percentages were collected for individual F–4 aircraft for the month of October 1967. Regressing these percentages against age and time since last

rework resulted in the following equation:

$$\text{(14)}\quad \underset{\text{(3.92)}}{1.7782} + 0.3539\ \text{Log}_e\ (\text{age}) + \underset{\text{(2.74)}}{0.207}\ \text{Log}_e\ (\text{time since last rework})$$

$$\text{Log}_e\ (\text{percentage of time not operationally ready})$$

$$\text{Correlation coefficient} = 0.30$$

$$\text{Standard error of estimate} = 0.7459$$

$$\text{Degrees of freedom} = 217.$$

Costing of material and man-hours. Because we cannot cost military man-hours or measure the material usage by individual aircraft, labor and material costs have been derived on the basis of the costs estimated at the Naval Air Rework Facility at North Island. The cost of labor and material has been estimated at \$14 per man-hour and is derived by dividing the average cost of a Progressive Aircraft Rework by the average man-hours expended. The cost of support man-hours is taken to be \$6.70 because virtually no material is consumed washing the aircraft, preparing it for flight, etc.

The imputed cost of downtime. The cost of an aircraft day will be taken as \$5,940. The imputed cost of downtime at the squadron for a quarter will therefore be:

$$\text{(15)}\quad \text{Imputed cost of downtime} = \frac{[e^{1.7782}(\text{age})^{0.3539}(\text{time since last rework})^{0.207}]}{100} \times [90\ \text{days}]\ [\$5940].$$

The equation that is used to estimate the maintenance costs in relation to age, t_1, and tour length, t_2, on a quarterly basis is:

$$\begin{aligned}\text{(16)}\quad \text{Maintenance cost} &= (\text{unscheduled man-hours})\ \$14 + (\text{support man-hours})\ \$6.70\\ &+ (\text{field team man-hours})\ \$14 + \text{imputed cost of downtime}\\ &+ [(\text{man-hours expended for calendar inspection})\ \$14\\ &+ (\text{days in process})\ \$5940] \text{ added in every 30 weeks.}\end{aligned}$$

$$\begin{aligned}\text{(17)}\quad U_N(t_1, t_2) &= (e^{6.5515}t_1^{0.0443}t_2^{0.1401})\ \$14 + (3\times 503)\ \$6.70 + (e^{4.0075}t_2^{0.2699})\ \$14\\ &+ \left(\frac{(e^{1.7782}t_1^{0.3539}t_2^{0.207})}{100}\right)(90\ \text{days})\ (\$5940) + ((805)\ \$14\\ &+ (8)\ \$5940) \text{ added in every 30 weeks.}\end{aligned}$$

The following assumptions will be made because the cost equations are logarithmic:

- $U_N(o, o) \equiv U_N(1, 1).$
- $U_N(t_1, o) \equiv U_N(t_1, 1).$

Figure 4 shows the maintenance costs as a function of age and time since last rework excluding the cost of a calendar inspection. The continuous curve in Figure 5 indicates the percent of time not operationally ready for an aircraft that has never been reworked. The discontinuous curve in Figure 5 shows the effect on the percent of time not operationally ready of periodic rework and the time since last rework.

The rework cost function in the dynamic programming formulation is:

$$U_N(t_1, o) + R_N(t_1\, t_2) + \frac{1}{(1+r)} f_{N+1}(t_1 + 1, 1).$$

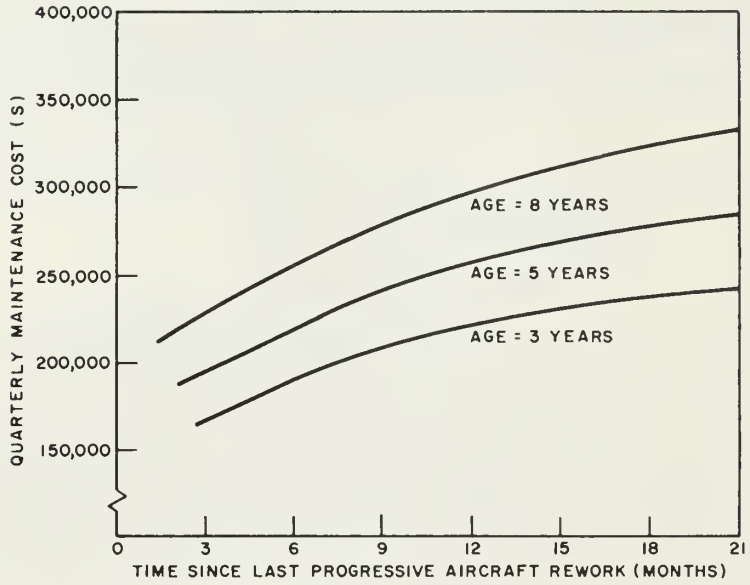


FIGURE 4. Estimated quarterly maintenance costs (including the imputed cost of downtime) as a function of time since last progressive rework and age.

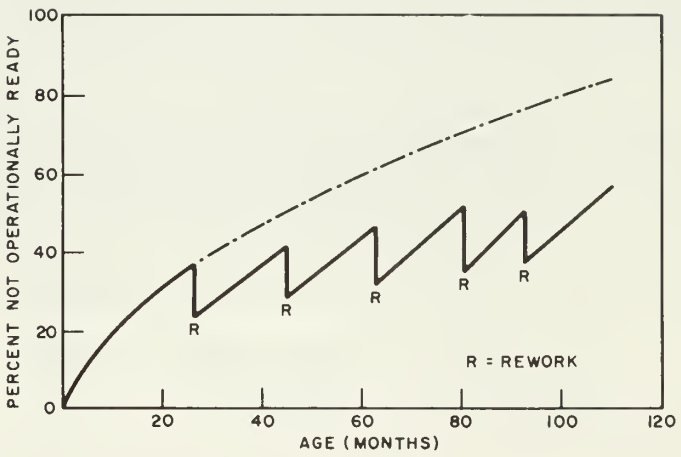


FIGURE 5. The effect of age and time since last rework on the percent not operationally ready.

The costs associated with a decision to rework are: the maintenance costs for the planning period, $U_N(t_1, o)$; the cost of the rework, $R_N(t_1, t_2)$; and the discounted future costs, $\frac{1}{(1+r)} f_{N+1}(t_1+1, 1)$. Because the aircraft is not being utilized for the entire planning period when a rework does occur, the total maintenance costs for the planning period should not be incurred. Therefore, only a portion of the maintenance costs, $U_N(t_1, o)$, will be allocated to the costs associated with a decision to rework. For a planning period of 90 days, $U_N(t_1, o)$ will be multiplied by

$$\frac{90-(\text{days in process for a progressive aireraft rework})}{90}.$$

VI. SOLUTION TO THE PROBLEM USING THE METHOD OF DYNAMIC PROGRAMMING

Replacement of the F–4A With the F–4J

For this problem, the residual value is held constant over the entire planning horizon. This is based on the assumption that the value of the components that could be salvaged from the F–4A will remain relatively stable for the length of time the F–4A remains in the inventory.

The residual value of the F–4A components, excluding engines, was estimated at $950,800 by Aviation Supply Office, Philadelphia, Pa. If the cost of two new engines is added to this figure, the residual value could be as high as $1,400,000. There will be a cost incurred disassembling the F–4A's and refurbishing their components. Therefore, $950,800 could be considered a lower limit and $1,400,000 an upper limit. An intermediate value of $1,100,000 will also be considered. The figure was derived by taking 80 percent of $1,400,000, based on the assumption by the Naval Air Rework Facilities that the cost of refurbishing components is approximately equal to 20 percent of their value.

Three computer runs were made using these residual values and the derived maintenance cost functions. The results are shown in Tables III, IV, and V. All runs began with the purchase of a new aircraft.

TABLE III

	Residual value = $950,800	
Tour	Decision	Length of tour (months)
I	Rework	27
II	Rework	18
III	Rework	18
IV	Rework	18
V	Rework	12
VI	Purchase	18
Purchase at the end of 9 years		

TABLE IV

Residual value = $1,100,000		
Tour	Decision	Length of tour (months)
I	Rework	27
II	Rework	18
III	Rework	18
IV	Purchase	27
Purchase at the end of 7.5 years		

TABLE V

Residual value = $1,400,000		
Tour	Decision	Length of tour (months)
I	Rework	27
II	Rework	18
III	Purchase	27
Purchase at the end of 6 years		

Another run was made using no residual value; for this case the purchase decision occurred at 11.7 years. This indicated that using residual values below $950,800 will not have a marked effect on the replacement age; the rate of increase in the replacement age appears to be very slow relative to a decrease in the residual value below $950,800.

The Optimal Rework Cycle

What appears to be extremely interesting and intuitively appealing is the relationship between tour length and aircraft age. The data showed that new aircraft require less maintenance and are operationally ready a larger percentage of the time during a tour relative to older aircraft. Therefore, the results indicated that new aircraft should have longer tours. Older aircraft are more expensive to maintain, require frequent major reworks, and have a lower availability than the newer aircraft. As the aircraft approaches the end of its service life, expensive repairs such as a Progressive Aircraft rework would not be undertaken as long as the aircraft is still safe for flight. The tour length would tend to increase before the aircraft is scrapped.

Sensitivity Analysis

The tour length and the replacement age vary considerably with the cost of an aircraft day and the percentage of time the aircraft is not operationally ready. A relatively slow rise in the percentage of time the aircraft is not operationally ready as a function of age and time since last rework will tend to delay reworks and purchase decisions. Decreasing the cost of an aircraft day below $5,940, the figure used in this study, will also have the same effect, i.e., stretch out tour lengths and delay purchases. For example, using a residual value of $1,100,000 and $5,000 for the cost of an aircraft day, the purchase decision occurred at 9.75 years; using $7,000 for the cost of an aircraft day forced a purchase at 6.75 years.*

Two computer runs were made, one without discounting future costs and one discounting them at 20 percent. Increasing the discount rate postpones the replacement decision, but does not effect the tour lengths. Undiscounted future costs decrease the replacement age of the aircraft and produce the same tour lengths as the computer runs using a 10- and 20-percent discount rate. For example, using a residual value of $1,100,000 and a 0 discount rate, the replacement decision while decreasing the discount rate decreases the replacement age of the aircraft.

*An iterative process should be used so that the cost of an aircraft day is compatible with the program service life. Convergence was tested and assured.

Summary and Conclusions

The results show that the replacement age is sensitive to the residual value of the aircraft that is being replaced, the cost of an aircraft day, and the percent of time the aircraft is not operationally ready at the squadron. For example, using a fast rising curve to describe the percent of time the aircraft is not operationally ready will tend to decrease tour lengths and decrease replacement age. Increasing the cost of an aircraft day above $5,940, the figure used in this study, will have the same effect. Increasing the discount rate postpones the replacement decision while decreasing the discount rate decreases the replacement age of the aircraft.

The analysis strongly indicates that older aircraft require more resources to maintain, are less available during a tour, and should be reworked more frequently than newer aircraft. This implies that it is unrealistic to plan for fixed tour lengths, to allocate resources to aircraft without considering their age and condition, and to assume that all aircraft are equally capable.

VII. AREAS FOR FURTHER STUDY

Other Problems That Could Be Solved With the Use of the Dynamic Program

The dynamic program could be used to determine whether a severely damaged aircraft should be restored or replaced. This could simply be accomplished by entering the program with the age and tour length of the damaged aircraft, and substituting the cost of restoring the aircraft as the cost of the next rework; the residual value would be the salvage value of the damaged aircraft. The keep decision would be suppressed because this would not be a feasible alternative.

If the aircraft were to be considered for use in the Naval Reserves or in a training squadron, the residual value would have to be adjusted to reflect the value of the aircraft in these secondary missions. The dynamic program would determine the age at which the aircraft would be retired from first-line status.

The Finite Planning Horizon

A critical assumption made was that the process stopped at the end of the 80th period; this assumes that all costs beyond the 80th period are zero. Carrying these zero future costs into the analysis will affect the decisions in the later stages of the program. How strong this effect is, and how far into the present it penetrates, are areas that need further investigation.

Appendix

AN EXAMPLE SHOWING THE EQUIVALENCE OF THE PRESENT VALUE SOLUTION WITH THE DYNAMIC PROGRAMMING SOLUTION

Suppose we are considering two alternatives, purchase (P) or keep (K), and a planning horizon of 3 years. The relevant costs are shown in the following matrices.

TABLE A–I. *Costs For Machine Made In Year 1*

Age	0	1	2
Maintenance	20	20	25
Replacement	200	220	240

TABLE A–II. *Costs For Machine Made In Year 2*

Age	0	1	
Maintenance	15	20	
Replacement	210	230	

TABLE A–III. *Costs For Machine Made In Year 3*

Age	0		
Maintenance	15		
Replacement	240		

Some examples clarify the use of the tables. If we enter the third period with a machine that is 2 years old, the operating and maintenance costs will be 25 (Table A–I). The cost of replacing a 1-year old machine at the beginning of the third period is 230 (Table A–II). The cost of a new machine in year 3 is 240 (Table A–III). Present value solution consists of evaluating all possible combinations of alternatives and then selecting the minimum cost combination. We begin the enumeration with the purchase of a new machine.

If the decision in year N is purchase, then the cost associated with this action is the maintenance costs of a new machine in year N plus the cost of replacing a machine in year N that is of age t. If the decision is keep (K), then the only cost in year N will be the maintenance costs of a machine of age t.

(A–1) $$(P_1,P_2,P_3)\ 220+235(1.10)^{-1}+245(1.10)^{-2}=636.1$$

(A–2) $$(P_1,P_2,K_3)\ 220+235(1.10)^{-1}+20(1.10)^{-2}=450.17$$

(A–3) $$(P_1,K_2,P_3)\ 220+20(1.10)^{-1}+255(1.10)^{-2}=448.9$$

(A–4 $$(P_1,K_2,K_3)\ 220+20(1.10)^{-1}+25(1.10)^{-2}=258.8$$

The minimum cost path is 4.

Dynamic Programming Solution

The recurrence relations for this problem are:

$$f_N(t)=\text{Minimum}\begin{cases} P\colon U_N(o)+C_N(t)+\dfrac{1}{(1+r)}f_{N+1}(1) \\ K\colon U_N(t)+\dfrac{1}{(1+r)}f_{N+1}(t+1), \end{cases}$$

where

$U_N(t)$ is the operating and maintenance cost of a machine in year N that is t years old,

$U_N(o)$ is the maintenance costs of a new machine purchased in year N,

$C_N(t)$ is the net cost of replacing a machine in year N that is t years old,

$f_N(t)$ is the cost at year N of the overall cost from a machine which is t years old, where an optimal replacement policy is employed for the remainder of the process,

N_o is the total number of periods that are being considered,

$f_{No+1}(t) \equiv 0$, and

r is the discount rate.

We begin the algorithm by evaluating all admissible values of the function $f_{N_o}(t)$ in the last period and then use these results to determine all admissible value of the function $f_{N_o-1}(t)$. This procedure continues through the first period where the minimum cost of the optimal policy is determined. Once the algorithm is completed, the optimal path can be traced out by following the minimum cost decisions beginning with the first period:

$$f_3(1) = \text{Minimum} \begin{bmatrix} P\colon U_3(o) + C_3(1) = 15 + 230 \\ K\colon U_3(1) \qquad\quad = 20 \end{bmatrix} = 20$$

$$f_3(2) = \text{Minimum} \begin{bmatrix} P\colon U_3(o) + C_3(2) = 15 + 240 \\ K\colon U_3(2) \qquad\quad = 25 \end{bmatrix} = 25$$

$$f_2(o) = \text{Minimum} \begin{bmatrix} P\colon U_2(o) + C_2(1) + f_3(1)\cdot\dfrac{1}{(1+r)} = 15 + 220 + 20[1.10]^{-1} \\ K\colon U_2(1) + f_3(2)\cdot\dfrac{1}{(1+r)} = 20 + 25[1.10]^{-1} \end{bmatrix} = 42.73$$

$$f_1(o) = \quad P\colon U_1(o) + C_1(o) + 42.7[1.10]^{-1} = 20 + 200 + 38.8 \ = 258.8$$

We purchase in year 1 and move into year 2 with a machine that is 1 year old. The minimum cost decision in year 2 is keep (K). We move into year 3 with a 2-year-old machine where the decision is keep (K). As a check we can add up the discounted cost from this policy in the following way:

Year	Policy	Cost ($f_N^* - f_{N+1}^*$)
1	P	216.07
2	K	17.73
3	K	25
		258.8

where f_N^* is the cost in year N following the optimal path.

BIBLIOGRAPHY

[1] Alchian, A., "Economic Replacement Policy," The RAND Corporation Rept. R–224 (Apr. 12, 1952).

[2] Bellman, Richard E., "Equipment Replacement Policy," J. Soc. Indust. Appl. Math., **3**, 133–136 (1955).

[3] Bellman, Richard E. and Stuart E. Dreyfus, *Applied Dynamic Programming* (Princeton University Press, 1962).

[4] Boness, A. James and Arnold N. Schwartz, "A Cost Benefit Analysis of Military Aircraft Replacement Policies," Nav. Res. Log. Quart. **16**, 237–257 (June 1969).

[5] Dreyfus, S., "A Generalized Equipment Study," J. Soc. Indust. Appl. Math., **8**, 425–435 (1960).

[6] Hadley, G., *Nonlinear and Dynamic Programming* (Addison-Wesley, 1964).

[7] Johnston, J., *Econometric Methods* (McGraw-Hill, 1960).

[8] Schwartz, Arnold N., LCDR J. A. Sheler, and CDR C. R. Cooper, "A Dynamic Programming Approach to the Optimization of Naval Aircraft Rework and Replacement Policies," Institute of Naval Studies, Study 20, Center for Naval Analyses.

[9] Suits, D. B., "Use of Dummy Variables in Regression Equations," J. Am. Stat. Assn., **52**, 548–551 (Dec. 1957).

A NOTE ON A FIRST APPLICATION OF CLUSTERING PROCEDURES TO FLEET MATERIAL CONDITION MEASUREMENTS

Henry Solomon

Program in Logistics

The George Washington University

ABSTRACT

The objective of this paper is to provide an independent evaluation of the nature and interpretation of ships' physical condition data generated by the USN Board of Inspection and Survey (INSURV). The substantive context is the classification of ships in terms of material condition and/or readiness based on scores pertaining to individual line elements within each ship. In order to account for multi-dimensional measures of each ship, clustering procedures are employed to evaluate existing ship classification systems and to indicate other possibilities.

1. INTRODUCTION

The creation and application of readiness and related measurements are rapidly becoming important areas of logistics research and management. While such measurements, if available, would have always been deemed useful, interest in obtaining such measurements has been accelerated. Two major motivations for this are increased attention given to cost-effectiveness analysis of logistics support systems and associated methodology for resource allocations, and command-control systems requiring information for strategic and tactical decisions involving deployment of forces.

It would be beyond the scope of this paper to attempt a comprehensive discussion of conceptual and methodological problems in deriving measurements of readiness of the fleet or any of its components. This paper has a much narrower scope. First, the context is limited to material, hence excluding such resources as personnel, supply, etc. Attention is further restricted to one information system pertaining to material, namely the information generated by the USN Board of Inspection and Survey (INSURV). The intent of this paper is an examination and evaluation of these data as to their nature and interpretation. This is pursued as a taxonomic or classification problem involving statistical data analysis and more specifically, clustering procedures. The substantive context is the classification of ships in terms of material condition and/or readiness.

2. INSURV PROCEDURES

There is a legal requirement that ships be periodically inspected with regard to their material condition. Accomplishing this inspection is the mission of INSURV. A team of experts spend about 2 or 3 days inspecting each ship, and the time interval between inspections for any one ship is in the order of 2–3 years.

For purposes of the inspection, each ship is characterized in terms of major "departments," a list of "functions" within each department, and a list of "material line elements" for each function. For a typical ship there are 12 departments, an average of five functions per department, and an average of approximately eight material line elements per function. Hence a typical ship has about 480 material

line elements which represent hardware entities. During an inspection, each material line element is assigned a grade intended to reflect its physical condition. Also an overall estimate is made of the ship's condition resulting in an assignment of a "satisfactory" or "unsatisfactory" grade. Exactly how this overall grade is determined is not known. It is likely some function of the physical condition grades of the material line elements and a subjective estimate of their importance. This may include as a special case, an estimate of all line elements being of equal importance to the ship.

In order to arrive at a measure of material condition of a ship, INSURV includes in its procedures a system of "weights" reflecting the relative importance of individual material line elements to the mission of the ship. These weights applied to the physical condition grades (where A, B, C, D, and E are treated as numerical values, 5, 4, 3, 2, and 1, respectively) of the respective line elements produce what is referred to as a ship's "Material Condition Indicator" (MCI). An example of the intent is to permit a very important line element with a poor physical condition grade to contribute differently to a ship's MCI than a line element of little importance with a poor physical condition grade. It should be noted that the calculation of a ship's MCI is accomplished independently of the assessment of a ship mentioned above as to its satisfactory or unsatisfactory condition.*

There are several important questions to be pursued with respect to the INSURV system. These include the relationship between the physical condition grades and the overall assessment of a ship's condition, the relationship between the weighted grades producing an MCI and the overall assessment of a ship, the relationship between an MCI and an unweighted indicator for each ship. These pertain to an important operational question which is the possibility of calculating and assigning a consistent and reliable overall assessment for each ship given only the grades or scores for individual line elements. This is of particular importance in considering readiness reporting systems outside of the INSURV environment. For example, it is important to obtain an estimate of the acceptability of ships reporting individual line element grades at frequent intervals and permitting the calculation of a ship's condition on a more current basis. Finally, there is the question of the classification system to be used for characterizing a ship's condition. For example, a system with two possible classifications (satisfactory or unsatisfactory) as compared to a system with three or more gradations.

3. ANALYSIS OF INSURV DATA

The remainder of this paper is an analysis of data for a sample of 123 INSURV inspections. The data analyzed do not include military value weights. Hence what follows does not pertain to questions involving MCI's. These will be explored in a later paper.

In order to evaluate the INSURV data, a sample of 123 inspections was taken as the data base. The criteria for choosing these inspections was that they pertain to the same type ships (i.e., destroyers) with no missing information. Again the data include only the physical condition grades and the overall assessments.

The 123 inspections included 65 satisfactory and 58 unsatisfactory ships. A first point of interest is the relationship between the average of all physical condition grades for each ship and the overall assessment. Assuming all line elements to be of equal or nearly equal value, it may be expected that ships with low grades would be declared unsatisfactory and those with high grades declared satisfactory. Table I shows the distribution of the number and cumulative percent of ships by grades and overall assessment.

*More detailed information concerning INSURV procedures may be found in [3], [4], and [5].

From Table I it may be observed that while there is hardly a direct correspondence between the ship's grades and their overall assessments, the unsatisfactory ships tend to have lower grades than the satisfactory ships. However, from the point of view of classifying a ship's condition given its average grade, some significant amount of inconsistency or "error" involving misclassification may occur.*

One major methodological problem is the representation of each ship by a single number whether it is the average physical condition grade or the MCI. The important alternative to this is the representation of a ship in a multidimensional manner or vector representation. Clustering is one valuable procedure in evaluating this alternative.

TABLE I. *Distribution of "Sat" and "Unsat" Ships By Physical Condition Grade*

Grade	Number of "Sat" Ships	Number of "Unsat" Ships	Cumulative Percent of "Sat" Ships	Cumulative Percent of "Unsat" Ships
2.7	2	1	3	2
2.8	2	1	6	3
2.9	1	13	8	26
3.0	10	15	23	52
3.1	12	6	42	62
3.2	8	5	54	71
3.3	7	2	65	74
3.4	5	2	72	78
3.5	4	5	78	86
3.6	5	2	86	90
3.7	5	4	94	97
3.8	3	2	98	100
3.9	1	0	100	100
Σ	65	58		

To obtain a relevant vector representation of each ship, the average physical condition grade for each of the 12 ship's departments was obtained. Summary measures of these grades are shown in Tables II–A and II–B. Table II–A pertains only to "Sat" ships portraying for each department the relative number of these ships by average physical condition grade. Table II–B portrays the same information for "Unsat" ships.

By comparing Tables II–A and II–B, it may be observed that the "Unsat" ships tend to have lower grades than "Sat" ships for each department; however, large proportions of "Sat" ships have low grades for various departments. Using summary data as shown in Tables II–A and II–B, it remains fairly difficult to determine or evaluate the consistency or existence of structural relationships between departmental grades and the overall ship assessments. To overcome this difficulty, clustering procedures have been used which are capable of efficiently handling vector representations.

*It has been observed that the same remark would apply to use of the MCI's as the MCI's are very close to the average physical condition grade. For the 123 inspections, the average absolute deviation is 0.09.

TABLE II–A. *Distribution of Grades Per Department For "Sat" Ships*

Grade	Department												
	1	2	3	4	5	6	7	8	9	10	11	12	Σ
5.0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
4.0–4.9	0.077	0.185	0.184	0.077	0.000	0.000	0.046	0.015	0.000	0.108	0.154	0.031	0.073
3.0–3.9	0.630	0.461	0.677	0.615	0.754	0.630	0.770	0.493	0.615	0.600	0.692	0.785	0.644
2.0–2.9	0.292	0.354	0.139	0.308	0.246	0.369	0.185	0.492	0.384	0.292	0.154	0.185	0.284
1.0–1.9	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

TABLE II–B. *Distribution of Grades Per Department For "Unsat" Ships*

Grade	Department												
	1	2	3	4	5	6	7	8	9	10	11	12	Σ
5.0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
4.0–4.9	0.069	0.052	0.052	0.052	0.000	0.000	0.000	0.017	0.000	0.086	0.086	0.052	0.039
3.0–3.9	0.586	0.396	0.638	0.535	0.707	0.586	0.534	0.379	0.397	0.448	0.569	0.569	0.529
2.0–2.9	0.345	0.551	0.310	0.414	0.293	0.414	0.448	0.603	0.603	0.466	0.345	0.379	0.431
1.0–1.9	0.000	0.000	0.000	0.000	0.000	0.000	0.017	0.000	0.000	0.000	0.000	0.000	0.001

4. APPLICATION OF CLUSTERING PROCEDURES

Alternative clustering procedures are available and are being applied to the INSURV data. The results described in this paper are limited to the application of a step-wise clustering procedure developed by Benjamin King [2].* Briefly, the King program starts by computing a correlation matrix. In the present context this is a 123 by 123 matrix involving the 123 inspections, where the correlations are in terms of the 12 departmental measurements for each inspection. Once the matrix is formed, all elements are scanned to determine the maximum coefficient. The two variables (i.e., inspections) which are represented by the maximum coefficient are summed creating a new variable or the first cluster or family. Once the first pass has been completed, the order of the matrix is reduced by one. At the second pass, the largest coefficient is determined in the $(n-1)\times(n-1)$ matrix. As a result of this pass, a third variable, i.e., inspection, may join the cluster formed as a result of the first pass, or join another inspection to form a new cluster. This process continues until the $n-1$st pass, in which all variables are clustered together forming one family.

Due to limitations of the computer immediately available for the program, the 123 inspections were treated in three parts involving the first 50 inspections, the second 50 inspections, and the remaining 23 inspections.

5. RESULTS OF CLUSTERING PROCEDURES

Included in Table III are partial results of clustering applied to the first 50 inspections (i.e., through the 10th pass). As a result of the first pass, two inspections clustered. Both of these were declared "Unsat." It should be noted that the overall assessment is not included in the measurements employed in the clustering procedure. The clusters are formed solely on the basis of vector representations of the ships as derived from the departmental grades. The result of the second pass is the formation of another cluster involving two other ships declared "Unsat," and so on.

*Other procedures being applied are the Fortier and Solomon method [1] and variants of this method which differ from King's method in that optimal clustering is obtained in terms of an objective function. However, these procedures are much more demanding in computer time.

TABLE III. *50 Inspections Sample*

Pass Number	Correlation coeff.	Cluster
1	0.9773	UU
2	0.9372	UU
3	0.9330	UU UU SS
4	0.9240	UU UUU SS
5	0.9043	UU UUU SS UU
6	0.8885	UU UUU UU SS UU
7	0.8727	UU UUU UU SU SS UU
8	0.8576	UU UUUU UU SU SS UU
9	0.8544	UU UUUU UUU SU SS UU
10	0.8299	UU UUUU UUU UUU SU SS

The important observation is that the inspections do cluster in a manner consistent with the overall assessments. This is not to suggest the existence of perfect homogeneity within all clusters. Such homogeneity is not to be expected.

The results of a complete set of passes for the 23-inspection sample may be seen in an abbreviated form in Table IV. Several observations may be made. First, the 23 inspections break into two completely dissimilar clusters as evidenced by the negative correlation coefficient as a result of the 22nd pass. The next observation is that the first of these clusters contains a mixture of seven "Unsat" and 13 "Sat" inspections. Given only these 23 inspections, if it is required that the number of classifications be restricted to two, at least one classification must remain fairly heterogeneous suggesting that two classifications may not be sufficient. It may be noticed that increasing the number of possible classifications to three, produces a much more acceptable system. For the data used in this illustration, a classification system of "Very satisfactory," "Satisfactory," and "Unsatisfactory," produces a well defined and relevant system.

TABLE IV. *23 Inspections Sample*

Pass number	Correlation Coeff.	Cluster
1	0.8560	SS
2	0.8161	SS SU
3	0.8080	SSS SU
4	0.7986	UU SSS SU
.	.	.
.	.	.
.	.	.
19	0.3175	UUSUUUSSUSS SSSSSSSS SSS
20	0.2396	UUSUUUSSUSSU SSSSSSSS SSS
21	0.1034	UUSUUUSSUSSUSSSSSSSS SSS
22	−0.0014	UUSUUUSSUSSUSSSSSSSSSSS

6. CONCLUSIONS

The results obtained thus far are sufficient to indicate via the above type of data analysis that there are structural relationships between physical condition grades associated with the various departments and the overall assessments. These assessments as data by themselves are not to be considered as

whimsical. Hence some analyses of these data which are employed, such as a time series of proportion of "Sat" and "Unsat" ships, may adequately describe the trend of the physical condition of the fleet.

A second conclusion is that from clustering procedures, meaningful classification systems may be created which contain a minimum number of meaningful classifications. The particular procedure described in this paper suggests strong possibilities in this direction. To achieve this objective, optimal clustering techniques such as described in [1] should be utilized. The results of the King procedure point the way to the feasibility of such an objective.

Finally, it must be noted that from a methodological point of view, a multi-dimensional or vector representation of a ship's condition provides a much more useful and meaningful measure than a scalar. Also clustering procedures provide a valuable tool for utilizing these vector representations in attempting to create and evaluate meaningful operational classification systems.

7. ACKNOWLEDGMENTS

The author is indebted to Professor Herbert Solomon of Stanford University for bringing the important area of clustering procedures to his attention. Messrs. Thomas A. Enger and William G. Underhill, previously with the Logistics Research Project, provided invaluable programming and computational assistance.

REFERENCES

[1] Fortier, J. J. and H. Solomon "Clustering Procedures," *Proceedings of the International Symposium on Multi-Variate Analysis* (Paruchuri P. Krishnaiah, ed.) (Academic Press, New York, 1966), pp. 493–506.

[2] King, B., "Step-Wise Clustering Procedures," J. Am. Statist. Assoc. **62**, 86–101 (1967).

[3] Marlow, W. H., S. J., Mathis, Jr. and I. S. Tolins, "A Framework for Evaluating System Readiness," Technical Memorandum Serial TM–21927. Logistics Research Project, The George Washington University (1968).

[4] Segal, F. W., "The Methodology for Improved Inspection Procedures by the U.S. Navy Board of Inspection and Survey," Technical Paper Serial T–160, Logistics Research Project, The George Washington University (1964).

[5] Tennant, Christine S. (1966). "INSURV Decision-Oriented Material Condition Indicator Model—Preliminary Analysis," Technical Memorandum Serial TM–13500, Logistics Research Project, The George Washington University (1966).

A NOTE ON ADAPTIVE BOILER TUBE PULLING

J. W. Devanney III

Massachusetts Institute of Technology

ABSTRACT

This paper develops an adaptive algorithm for determining boiler tube pulling strategies by postulating a Beta prior on the probability that an individual tube is defective. This prior is updated according to Bayes' Rule as a result of the sample obtained during the tube pulling process.

INTRODUCTION

The biggest single job in thc overhaul of steam turbine ships is usually the inspection and repair of boiler tubes. In general, not all and sometimes very few of the 1,200 or so tubes in a boiler need replacement; however, one cannot determine for certain whether a tube needs replacement without going through the laborious and expensive process of pulling it. On the other hand, failure to replace a defective tube can lead to substantial loss of power and expensive shutdowns, perhaps at critical moments in the ship's career. Finally, the best information one has on the likelihood of the various condition of the tubes in the boiler is the condition of the tubes already pulled. The problem thus arises: how many tubes should we pull and how should we change this decision as the state of the pulled tubes becomes known to us?

Consider the following idealized boiler. The boiler has N tubes. A tube is either defective or not defective—there are no in-betweens. More discriminating models are possible, which distinguish between say, a failed tube, a badly corroded tube, a lightly corroded tube, and a perfect tube at some expense in computational feasibility. For now, we adhere to the binary case. We will assume further that defective tubes are generated by a Bernoulli process with unknown parameter, p, which we will attempt to estimate from the condition of the tubes already pulled. Roughly, if many of the tubes are bad, then it is more likely that p is close to 1.00 and thus more likely that the next tube(s) to be pulled are defective.

The assumption of a single Bernoulli process assumes that there are no locational differences in the rates of corrosion. This is unrealistic for different parts of the boiler experience very different fire-side and steam conditions, and, in fact, different corrosive mechanisms apply to different parts of the boiler. The model can be extended to cover some of these differences by regarding the superheater, steam generator, economizer as each a separate 'boiler' with its own p.

Let $C(m)$ be the cost associated with pulling and replacing m tubes in a block. This should be the marginal cost, given that the boiler is open and under repair. In general, $C(m)$ will not be linear in m reflecting economies due to pulling tubes in a block. Let $F(k)$ be the expected cost of not fixing k tubes which are defective. These costs should include the cost of downtime, of plugging, and of the resultant loss in power.

DERIVATION OF RELEVANT PROBABILITIES

We will assume that the occurrence of defective tubes is governed by a Bernoulli process with parameter, p, where p is the probability of a defective tube on a single trial. Under this assumption and given p if we pull m tubes in a block the distribution of k, the number of defective tubes, is the familiar binomial

$$f_B(k|p, m) = \frac{m!}{k!(m-k)!} p^k (1-p)^{m-k}. \tag{1}$$

Unfortunately, we don't know what p is, and further, as the results of our earlier investigation of the boiler become clear presumably our ideas about p will change.

The Bayesian takes the view that p, being an unknown, is a random variable, and like any random variable it has a distribution. What can we say about this distribution? Well, since p is a probability it must be between 0.0 and 1.0, inclusive. Let us assume that before pulling any tubes, we have no reason to believe that p is any more likely to be any one number between 0 and 1 than any other. In this case our state of (un)knowledge about p is described by a uniform density function,

$$f(p) = 1 \qquad 0 \leq p \leq 1.$$

An analytic representation of this density function can be obtained by using the Beta density function, $f_\beta(p|r', n')$, where the parameters r' and n' equal 1 and 2, respectively.

In general, the family of Beta density functions is given by

$$f_\beta(p\,|\,r,n) = \frac{1}{B(r,n)} p^{r-1}(1-p)^{n-r-1} \qquad \begin{matrix} p \leq 1 \\ r,n > 0, \end{matrix}$$

where $B(r, n)$ is the complete Beta integral. The Beta density functions are positive only between 0 and 1, and by varying the parameters r and n a quite rich family of distributions can be obtained. Almost any smooth unimodal and some bimodal density functions on the interval [0, 1] can be closely approximated by a Beta density. Given the ability to approximate a wide range of possible distributions on the unknown likelihood of a defective tube, p, we will lose little generality if we decide to limit ourselves to Beta densities in choosing a distribution on p. This essentially harmless limitation has some very significant analytical advantages. First, given that we do have a Beta distribution on p, $f_\beta(p|r,n)$ the probability of k defective tubes in the next m trials is given by:

$$Pr\ \begin{matrix}(k \text{ bad tubes out of } m \text{ given Beta} \\ \text{density on } p \text{ with parameters } r \text{ and } n)\end{matrix} = \int_0^1 f_B(k|p,m) \cdot f_\beta(p|r,n)\,dp \tag{2}$$

which expression simply says that the probability of k bad tubes in m trials is the probability of k bad tubes in m trials given p times the probability that the unknown parameter of the Bernoulli process equals p, summed over all possible p. By performing the integration, it can be shown that this probability is equal to

$$\frac{(k+r-1)!(n+m-k-r-1)!m!(n-1)!}{k!(r-1)!(m-k)!(n-r-1)!(m+n-1)!}$$

which distribution will be denoted by $f_{\beta B}(k\,|\,r,n,m)$ and is called the Beta binomial [1].

Secondly, if we assume a Beta density on p, the updating of this distribution on p to take into account our sample data reduces to an extremely simple algebraic manipulation. Suppose we arbitrarily start out with a uniform distribution—one value of p is as likely as any other value. Suppose further that we pull 20 tubes and observe that 5 are defective. The probability distribution on p after we have observed 5 defective tubes in 20 trials, which will be denoted by $f(p|5,20,r,n)$, is given by Bayes' Rule which in this case takes the form

$$f(p|5,20,1,2) = \frac{f_B(5|p,20)\, f_\beta(p|1,2)}{\int_0^1 f_B(5|p,20)\, f_\beta(p|1,2)\,dp} \tag{3}$$

Making the substitutions on the right-hand side, we have

$$f(p|5,20,1,2) = \frac{\binom{20}{5} p^5(1-p)^{20-5}(B(1,2))^{-1}p^{1-1}(1-p)^{2-1-1}}{\int_0^1 \binom{20}{5} p^5(1-p)^{20-5}(B(1,2))^{-1}p^{1-1}(1-p)^{2-1-1}dp} \tag{4}$$

$$= \frac{p^{(5+1)-1}(1-p)^{(20+2)-(5+1)-1}}{B(5+1,20+2)}$$

$$= f_\beta(p|5+1,20+2).$$

Thus, our new distribution on p is also a Beta whose parameters r'' and n'' are given by

$$r'' = 5+1 = 6 \qquad \text{(more generally) } r'' = r + r'$$

and

$$n'' = 20+2 = 22 \qquad \text{(more generally) } n'' = n + n',$$

where n and r are now the number of tubes already pulled and the number of these tubes which have proved defective. This new distribution on p, $f_\beta(p|6,22)$, is shown in Figure 1. Note, as might be expected, we now more heavily weight p's near 1/4 and also our new distribution is considerably tighter than the old, reflecting the increase in our knowledge about p. Generalizing, at any time in the process, after we have observed r bad tubes out of a total of n pulled, our distribution on p is given by

$$f_\beta(p|r+r',n+n'),$$

where r' and n' are the parameters of the original distribution. It should be clear that as long as r' and n' are small (i.e., the original distribution is wishy-washy) it will not be long before $r'' \simeq r$, and $n'' \simeq n$. That is, the data will overwhelm our original feelings. Hence, the seeming obstacle of a choice of original distribution is no real problem at all, as long as we don't have any strong feelings about p before we start pulling.*

*If we do have strong feelings (say, we are sure p is near 1), we should pick an original distribution which is consistent with these feelings and our results will be affected accordingly. In general, the prior will depend on a preliminary inspection of the tube bank and the general condition of the boiler.

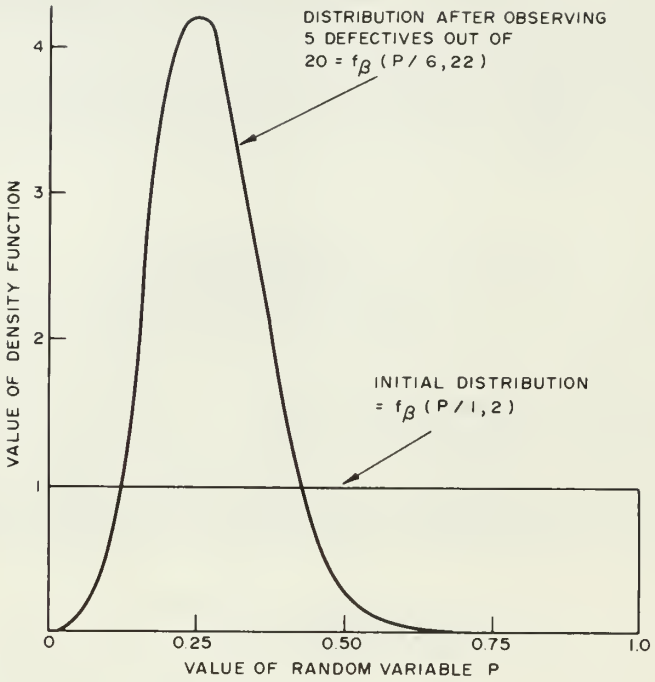


FIGURE 1. Distributions on p

Given this background and generalizing the above argument leading to the Beta binomial, the probability of observing k defective tubes out of a batch of m pulled after we have already pulled n tubes and r of these were defective is simply the Beta binomial whose parameters are the parameters of the present distribution on p, and m:

$$f_{\beta B}(k|r+r', n+n', m).$$

AN ALGORITHM FOR DETERMINING THE OPTIMAL TUBE PULLING POLICY, GIVEN THE FOREGOING PROBABILISTIC CONSIDERATIONS

We now proceed to the development of a dynamic program for determining minimal expected cost tube pulling strategies—strategies which, since they realize our probabilities will change as we move through the process, properly weight the value of experimentation.

At any point in the tube pulling process, the current situation can be described by the pair (n, r) where n is the number of tubes already pulled and r is the number of these tubes which have turned out to be defective. We define a function $W_n(r)$ over all possible combinations of n and r to be the minimum attainable expected cost associated with any further tube pulling and any defective tubes left unrepaired, given that we have already pulled n tubes, r of which were defective. We proceed to develop a recursive relationship for $W_n(r)$. At any point (n, r) in the process, we have the following alternatives:

1. We can stop pulling tubes, close up the boiler, and bear the expected costs associated with any defective tubes we have failed to repair.

2. We can pull m tubes, $0 < m \leq N-n$, and pay the cost, $C(m)$, associated with this job and then see where we are.

Let us develop the minimum expected costs associated with each of these alternatives. Assuming there are N tubes in the boiler and we have already pulled n of these and have observed r defective

ones, by the foregoing arguments the probability that k of the remaining $N-n$ tubes are bad is $f_{\beta B}(k|r+r', n+n', N-n)$. Thus, the expected future costs of not pulling any more tubes after having observed r defective out of n is:

$$\sum_{k=0}^{N-n} f_{\beta B}(k|r+r', n+n', N-n)F(k),$$

where $F(k)$ is the expected cost associated with k nonrepaired defective tubes.

If after having observed r bad tubes out of n pulled we decide to keep pulling, the situation is a little more complicated. Suppose we decide to pull m tubes and see where we are. Well, we will certainly bear the costs of pulling and replacing m tubes, $C(m)$. Out of the m tubes pulled, k will be defective where k is a random variable between 0 and m inclusive. By the earlier argument the density function on k is $f_{\beta B}(k|r+r', n+n', m)$. After we have pulled m tubes and k of them turned out to be defective, we will be faced with the original choice only now we will have pulled $n+m$ tubes and will have observed $r+k$ defective. But the minimum expected cost attainable from this latter situation forward is by definition $W_{n+m}(r+k)$. In short, the minimum expected $n+m$ cost associated with the alternative of pulling $m>0$ tubes after pulling n and observing r defective is

$$C(m)+\sum_{k=0}^{m} f_{\beta B}(k|r+r', n+n', m)\cdot W_{n+m}(k).$$

But after having pulled n tubes and observed r failures, we will want to follow the expected cost minimizing alternative. Moreover, the value of that alternative will be $W_n(r)$ or

$$W_n(r)=\underset{0\le m\le N-n}{\text{minimum}}\begin{cases}\displaystyle\sum_{k=0}^{M-n} f_{\beta B}(k|r+r', n+n', N-n)\cdot F(k) & \text{if } m=0\\ C(m)+\displaystyle\sum_{k=0}^{m} f_{\beta B}(k|r+r', n+n', m)\cdot W_{n+m}(k) & \text{if } m>0.\end{cases}$$

This expression holds for all n, $0\le n<N$ and all r, $0\le r<n$.* $W_n(r)$ can be tabulated by observing that if we keep pulling tubes, sooner or later we will get to $n=N$. But after having pulled all the tubes, the sum of the costs of future tube pulling and nonrepaired tubes is 0. For example, $W_N(r)=0$ for all r. Thus, one substitutes $W_N(r)=0$ into the right-hand side of the above equation, which expression yields $W_{N-1}(r)$ for all r. $W_N(r)$ and $W_{N-1}(r)$ can then be used in the right-hand side, to calculate $W_{N-2}(r)$ for all r. Proceeding backwards, we can recursively calculate $W_n(r)$ for all possible combinations of n and r. While doing this one keeps track of the expected cost minimizing choice (number of tubes pulled) for each combination of n and r. Call this table $M_n(r)$. $M_n(r)$ then forms the optimal policy, indicating to the foreman how many tubes should be pulled in the next block, given any possible situation he might get into. A *PL/1* program implementing this algorithm has been written, and the routine is currently being considered for implementation by the Atlantic Cruiser-Destroyer Fleet.

REFERENCES

[1] Raiffa, H. and R. Schlaifer, *Applied Statistical Decision Theory* (Boston, 1961), p. 237.

*Dynamic programmers will note that this recursion has the interesting property that the ensuing stage as well as the state is a function of the current decision.

A NOTE ON A STOCHASTIC PRODUCTION-MAXIMIZING TRANSPORTATION PROBLEM

Uri Yechiali

New York University

ABSTRACT

A stochastic production-maximizing problem with transportation constraints is considered where the production rates, R_{ij}, of man i — job j combinations are random variables rather than constants. It is shown that for the family of Weibull distributions (of which the Exponential is a special case) with scale parameters λ_{ij} and shape parameter β, the plan that maximizes the expected rate of the entire line is obtained by solving a deterministic fixed charge transportation problem with no linear costs and with "set-up" cost matrix $||\lambda_{ij}||$.

The Time-Minimizing Transportation Problem (TMTP) was treated by Barsov (1959) [2] and again by Hammer (1969) [4] and may be stated as follows: Given a set of m origins and n destinations, where there are a_i $(i=1,2, \ldots, m)$ units available at the ith origin and b_j $(j=1,2, \ldots, n)$ units required at the jth destination (such that, $\sum_{i=1}^{m} a_i = \sum_{j=1}^{n} b_j$), find a set of nonnegative variables x_{ij} $(i=1,2, \ldots, m;$ $j=1,2, \ldots, n)$ satisfying the (classical transportation-type) constraints

$$\sum_{j=1}^{n} x_{ij} = a_i \qquad (i=1,2, \ldots, m)$$

$$\sum_{i=1}^{m} x_{ij} = b_j \qquad (j=1,2, \ldots, n) \tag{1}$$

and minimizing the *greatest* of the given nonnegative numbers t_{ij} for which $x_{ij} > 0$.

The t_{ij} may be interpreted as the time required to transport a positive load x_{ij} (however big or small) from the ith origin to the jth destination. Thus the problem is to find a transportation plan which makes the most time-consuming trip as short as possible.

In a production context the analogous problem will be to consider the *rate* of production, R_{ij}, instead of the time t_{ij}, and to seek a production plan $X=\{x_{ij}\}$ satisfying (1) and maximizing the *smallest* of the given nonnegative rates R_{ij} for which $x_{ij} > 0$. The R_{ij}'s are now interpreted as the rate of production (on a production line, say) of a man belonging to group (origin) i when he is assigned to job (destination) j. As above, there are a_i men available in the ith group and b_j men required for the jth job.

For any production plan X that satisfies (1) let $A_X=\{(ij)|x_{ij}>0\}$. For any such plan, the rate of production of the entire line, R, will be given by

$$R=\underset{(ij)\in A_X}{\text{Minimum}}(R_{ij}) \tag{2}$$

and the problem is then to find a plan for which R is as large as possible.

Now, suppose that for each man i—job j combination, the corresponding R_{ij} is not a constant, but a continuous nonnegative random variable with distribution function $F_{ij}(\cdot)$. This implies that, for any plan X, R (as given by (2)) is also a random variable. Our objective then is to find a production plan that will maximize the *expected* rate of production of the line, i.e., we seek a plan X satisfying (1) so as to achieve

$$\max_X\{E[R]\}=\max_X\{E[\underset{(ij)\in A_X}{\text{Minimum}}(R_{ij})]\}. \tag{3}$$

Assuming that the R_{ij}'s are independent random variables with finite means, the distribution function of R, $F_R(\cdot)$, is found to be

$$F_R(r)=1-\prod_{(ij)\in A_X}[1-F_{ij}(r)],$$

and the expected rate of production is given by

$$E[R]=\int_0^\infty\left\{\prod_{(ij)\in A_X}[1-F_{ij}(r)]\right\}dr.$$

Now consider the family of Weibull distributions where R_{ij} has a scale parameter $\lambda_{ij}>0$ and shape parameter $\beta>0$ (equal for all man-job combinations). In this case, the distribution function of R_{ij} is

$$F_{ij}(r)=1-\exp(-\lambda_{ij}r^\beta),\qquad r\geq 0. \tag{4}$$

We consider also the following "Fixed Charge Transportation Problem" (FCTP) with no linear costs: Given $\lambda_{ij}>0$ $(i=1,2,\ldots,m;j=1,2,\ldots,n)$ find a plan X satisfying (1) so as to achieve

$$\text{Min}\left\{\sum_{(ij)\in A_X}\lambda_{ij}\right\}. \tag{5}$$

We now show the following:

THEOREM: The solution to the stochastic production-maximizing transportation problem (3) is given by the solution of the FCTP (5).

PROOF: For any given plan X we obtain:

$$E[R] = \int_0^\infty \exp\left\{-\left(\sum_{(ij)\in A_X} \lambda_{ij}\right) r^\beta\right\} dr \tag{6}$$

$$= \left(\frac{1}{\sum_{(ij)\in A_X} \lambda_{ij}}\right)^{1/\beta} \Gamma\left(\frac{1}{\beta}+1\right),$$

where $\Gamma(\cdot)$ denotes the Gamma function. It is clear that $E[R]$ in (6) is maximized whenever $\sum_{(ij)\in A_X} \lambda_{ij}$ is minimized. This completes the proof.

By letting $\beta = 1$ in (4) it is readily seen that the exponential family of distributions is a special case of the family of Weibull distributions. Note also that if we let $m = n$ and $a_i = b_j = 1$ for all i and j then the deterministic and stochastic production-maximizing problems are transformed, respectively, into the classical [3] and stochastic [6] bottleneck assignment problems, whereas the FCTP [1] is transformed into the assignment problem.

In general, fixed charge problems have proven difficult to solve, primarily because each extreme point (here a basic solution of (1)) of the convex set of feasible solutions is a local optima. In our case, however, a direct way to solve the FCTP would be to enlarge it into an assignment problem of order

$$\left(\sum_{i=1}^{m} a_i\right) \times \left(\sum_{j=1}^{n} b_j\right).$$

Another approach could be to formulate the FCTP as an all-integer linear program [1]. A third method would employ a branch and bound algorithm as presented in [5]; however, for large problems all of the above methods would eventually become inefficient, and an approximative procedure (such as the one suggested in [1]) seems to be more practical. Additional references for approximative methods may be found in [5].

In summary, we have shown that the plan that maximizes the expected production rate of the entire line in a randomized production-maximizing transportation problem can be found by solving a deterministic fixed charge transportation problem with no linear costs and with fixed-charge cost matrix $\|\lambda_{ij}\|$ whose entries are the scale parameters of the random variables R_{ij}.

REFERENCES

[1] Balinski, M. L., "Fixed-Cost Transportation Problems," Nav. Res. Log. Quart. **8**, 41–54 (1961).

[2] Barsov, A. S., *What is Linear Programming* (D.C. Heath and Co., Boston, 1964), (translated from the Russian edition, 1959).

[3] Gross, O., "The Bottleneck Assignment Problem," the RAND Corporation, P–1630 (Mar. 6, 1959).

[4] Hammer, P. L., "Time-Minimizing Transportation Problems," Nav. Res. Log. Quart. **16**, 345–357 (1969).

[5] Steinberg, D. I., "The Fixed Charge Problem," Nav. Res. Log. Quart. **17**, 217–235 (1970).

[6] Yechiali, U., "A Stochastic Bottleneck Assignment Problem," Manag. Sci. **11**, 732–734 (1968).

NATO CONFERENCE: APPLICATION OF OPERATIONAL RESEARCH TO TRANSPORT PROBLEMS, August 14–18, 1972

The Conference, sponsored by the NATO Advisory Panel on Operations Research as part of its program for 1972, will be held in Sandefjord, Norway. Attendance will be limited to 120 persons. Attendees are expected to arrange for their own financial support.

The Conference will explore the uses of operational research in studying transport problems in both the military and civilian sectors. The scientific program will include formal sessions in the following major subject areas: assessment of transport technological advances, techniques of analysis oriented to transport problems, identification of major transport problem areas, interplay between civilian and military transport activities, the role of computers in transport management and planning, and selection and evaluation of transport infrastructures. There will also be panel sessions concerning future operational research needs in transport, role of operational researchers in a multimodal situation, and perspectives on transport operational research society activities and relationships.

Abstracts of papers or panel presentations are due no later than December 15, 1971, and should be addressed to the Scientific Director: Dr. Murray A. Geisler, The Rand Corporation, 1700 Main Street, Santa Monica, California 90406.

Information Bulletins containing detailed information on participation and attendance at the conference may be obtained from the Scientific Director or the American Point of Contact: Mr. Russell F. Stryker, OASD (I&L) TD, Room 3C838, The Pentagon, Washington, D.C. 20301.

☆ U.S. GOVERNMENT PRINTING OFFICE : 1971 OL—438-870

INFORMATION FOR CONTRIBUTORS

The NAVAL RESEARCH LOGISTICS QUARTERLY is devoted to the dissemination of scientific information in logistics and will publish research and expository papers, including those in certain areas of mathematics, statistics, and economics, relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Manuscripts and other items for publication should be sent to The Managing Editor, NAVAL RESEARCH LOGISTICS QUARTERLY, Office of Naval Research, Arlington, Va. 22217. Each manuscript which is considered to be suitable material for the QUARTERLY is sent to one or more referees.

Manuscripts submitted for publication should be typewritten, double-spaced, and the author should retain a copy. Refereeing may be expedited if an extra copy of the manuscript is submitted with the original.

A short abstract (not over 400 words) should accompany each manuscript. This will appear at the head of the published paper in the QUARTERLY.

There is no authorization for compensation to authors for papers which have been accepted for publication. Authors will receive 250 reprints of their published papers.

Readers are invited to submit to the Managing Editor items of general interest in the field of logistics, for possible publication in the NEWS AND MEMORANDA or NOTES sections of the QUARTERLY.

NAVAL RESEARCH
LOGISTICS
QUARTERLY

SEPTEMBER 197
VOL. 18, NO. 3
NAVSO P-1278

CONTENTS

OFFICE OF NAVAL RESEARCH
Arlington, Va. 22217